Available online at www.sciencedirect.com
ScienceDirect
Bioorganic & Medicinal Chemistry 15 (2007) 1483–1503
Bioorganic & Medicinal Chemistry

TOMOCOMD-CARDD descriptors-based virtual screening of tyrosinase inhibitors: Evaluation of different classification model combinations using bond-based linear indices

Gerardo M. Casañola-Martín,[a,b] Yovani Marrero-Ponce,[a,c,*] Mahmud Tareq Hassan Khan,[d,e] Arjumand Ather,[f] Sadia Sultan,[g] Francisco Torrens[c] and Richard Rotondo[h]

[a]*Unit of Computer-Aided Molecular "Biosilico" Discovery and Bioinformatic Research (CAMD-BIR Unit), Department of Pharmacy, Faculty of Chemistry—Pharmacy and Department of Drug Design, Chemical Bioactive Center, Central University of Las Villas, Santa Clara, 54830 Villa Clara, Cuba*
[b]*Department of Biological Sciences, Faculty of Agricultural Sciences, University of Ciego de Avila, 69450 Ciego de Avila, 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]*Pharmacology Research Laboratory, Faculty of Pharmaceutical Sciences, University of Science and Technology, Chittagong, Bangladesh*
[e]*Department of Pharmacology, Institute of Medical Biology, University of Tromso, Tromso 9037, Norway*
[f]*The Norwegian Structural Biology Centre (NorStruct), University of Tromso, Tromso 9037, Norway*
[g]*HEJ Research Institute of Chemistry, University of Karachi, Pakistan*
[h]*Mediscovery, Inc. Suite 1050, 601 Carlson Parkway, Minnetonka, MN 55305, USA*

Received 4 October 2006; accepted 31 October 2006
Available online 2 November 2006

Abstract—A new set of bond-level molecular descriptors (bond-based linear indices) are used here in QSAR (quantitative structure–activity relationship) studies of tyrosinase inhibitors, for finding functions that discriminate between the tyrosinase inhibitor compounds and inactive ones. A database of 246 compounds was collected for this study; all organic chemicals were reported as tyrosinase inhibitors; they had great structural diversity. This dataset can be considered as a helpful tool, not only for theoretical chemists but also for other researchers in this area. The set used as inactive has 412 drugs with other clinical uses.
Twelve LDA-based QSAR models were obtained, the first six using the non-stochastic total and local bond-based linear indices as well as the last six ones, the stochastic molecular descriptors. The best two discriminant models computed using the non-stochastic and stochastic molecular descriptors (Eqs. 7 and 13, respectively) had globally good classifications of 98.95% and 89.75% in the training set, with high Matthews correlation coefficients (*C*) of 0.98 and 0.78. The external prediction sets had accuracies of 98.89% and 89.44%, and (*C*) values of 0.98 and 0.78, for models 7 and 13, respectively. A virtual screening of compounds reported in the literature with such activity was carried out, to prove the ability of present models to search for tyrosinase inhibitors, not included in the training or test set. At the end, the fitted discriminant functions were used in the selection/identification of new ethylsteroids isolated from herbal plants, looking for tyrosinase inhibitory activity. A good behavior is shown between the theoretical and experimental results on mushroom tyrosinase enzyme. It might be highlighted that all the compounds showed values under 10 μM and that **ES2** (IC_{50} = 1.25 μM) showed higher activity in the inhibition against the enzyme than reference compounds kojic acid (IC_{50} = 16.67 μM) and L-mimosine (IC_{50} = 3.68 μM). In addition, a comparison with other established methods was carried to prove the adequate discriminatory performance of the molecular descriptors used here. The present algorithm provided useful clues that can be used to speed up in the identification of new tyrosinase inhibitor compounds.
© 2006 Elsevier Ltd. All rights reserved.

Keywords: TOMOCOMD-CARDD software; Non-stochastic and stochastic bond-based linear indices; LDA-based QSAR model; Tyrosinase inhibitor; Ethylsteroid compounds; Ligand-based virtual screening.
* Corresponding author. Tel.: +53 42 281192/+53 42 281473/+34 963543156; fax: +53 42 281130/+53 42 281455/+34 963543156; e-mail addresses: ymarrero77@yahoo.es; yovani.marrero@uv.es; ymponce@gmail.com; yovanimp@qf.uclv.edu.cu
URL: http://www.uv.es/yoma/

0968-0896/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2006.10.067

1. Introduction

Melanin production is principally responsible for skin color and plays an important role in prevention of sun-induced skin injury. Melanin is produced by melanocytes in the basal layer of epidermis.[1] Melanocytes have specialized lysosome-like organelles, termed melanosomes, which contain several enzymes that mediate the production of melanin.[2] Enzyme tyrosinase (EC 1.14.18.1) is involved in this process. This copper-containing enzyme is widely distributed in nature. It catalyzes two different reactions using molecular oxygen: the oxidation of amino-acid tyrosine into DOPA (3,4-dihydroxy-L-phenylalanine) (monophenolase activity) and subsequently into DOPA-quinone (diphenolase activity) the rate-limiting step for the melanin biosynthesis.[3]

Various dermatologic disorders result in the accumulation of excessive levels of melanin in the epidermal pigmentation. These hyperpigmented lentigenes include melasma, age spots or liver spots, and sites of actinic damage (i.e., due to solar ultraviolet irradiation).[4]

Unfortunately, several purportedly active agents (e.g., arbutin and kojic acid, among others) have not been demonstrated yet to be clinically efficacious when critically analyzed in carefully controlled studies, pharmaceutical products containing 2–4% hydroquinone (HQ) are moderately efficacious, but HQ is considered to be cytotoxic to melanocytes and potentially mutagenic to mammalian cells.[4–6]

Therefore, the current therapies are considered inadequate for these conditions. Although many compounds have been reported as tyrosinase inhibitors,[7–10] their activities are not potent enough or harmful adverse effects are shown. Taking into account this consideration, the search for new natural products and synthetic compounds with such activity still continues.[11–13]

In this sense, one of our group's researches has been focused on finding new potent tyrosinase inhibitors through 'trial-and-error' techniques; recently, Khan et al. reported 2,5-disubstituted-1,3,4-oxadiazole analogues[14] exhibiting strong inhibitory activity against the enzyme. In another publication, the same research group has reported that (+)-androst-4-ene-3,17-dione as well as its five metabolic analogues having steroidal skeletons, namely androsta-1,4-diene-3,17-dione, 17β-hydroxyandrosta-1,4-dien-3-one, 11α-hydroxyandrost-4-ene-3,17-dione, 11α,17β-dihydroxyandrost-4-en-3-one, and 15α-hydroxyandrosta-1,4-dien-17-one, exhibited moderate inhibitory activities against the tyrosinase.[15] In 2004, Ahmad et al.[16] reported that a new coumarinolignoid 8′-*epi*-cleomiscosin **A** together with the new glycoside 8-*O*-β-D-glucopyranosyl-6-hydroxy-2-methyl-4*H*-1-benzopyrane-4-one exhibited strong inhibition against the tyrosinase enzyme, when compared to the standard tyrosinase inhibitors kojic acid and L-mimosine. The new coumarinolignoid exhibited twice more potency than that of the standard potent inhibitor L-mimosine.[16]

On the other hand, the usually expensive and time-consuming experimental tests (based on 'trial and error' screening), specially pharmacological and toxicological tests in developing new drugs, coupled with candidate attrition rates during the discovery and development processes, highlight the need for a 'sea change' in the drug discovery paradigm.[17]

To reduce costs, pharmaceutical companies have to find new technologies to replace the old 'hand-crafted' synthesis and test new chemical entities (NCE) approaches.[18] In this sense, cheminformatics can be used to analyze data from high-throughput screening (HTS) and other forms of chemistry, thereby aiding in the identification of optimal lead structures.[19]

There are two main ways in virtual screening techniques according to their particular modeling of molecular recognition and the type of algorithm used in database searching.[18,19] The principle of similarity (ligand-based methods)—similar compounds are assumed to produce similar effects[20]—and the principle of complementarity (receptor-based methods)–the receptor of a biologically active compound is complementary to the compound itself (i.e., a lock-and-key model). The selection of the method depends on the knowledge of the active molecules and their receptor.

Structure-based drug design is an approach routinely used in medicinal chemistry. This receptor-based method needs the NMR or X-ray crystallography structure of the macromolecule, in this case, the tyrosinase enzyme, which is still not discovered. However, use of 3D (three-dimensional) structural information suffers from the fact that the binding affinity for the ligands cannot be predicted with high degree of accuracy with the presently available methods.[21] Therefore, considering that many tyrosinase inhibitor compounds are known, ligand-based methods founded on QSAR (quantitative structure–activity relationships) models can be applied, in order to describe the biological activity of tyrosinase inhibitors.

In addition, when computational approaches based on discrimination functions are used, it is possible to classify active chemicals from inactive ones, and predict beforehand the biological activity of new lead compounds with better therapeutic profiles, providing a useful tool to solve the ancestral problem of the '*trial and error*' methods for selecting a compound with a desired property.

Therefore, cheminformatical in silico methods are playing an important role in the drug discovery paradigm, and have the potential for transforming early stage drug discovery, particularly in terms of time and cost savings.[22] In relation to it, several authors report a high incidence of the use of novel molecular descriptors to develop QSAR studies of in silico drug screening.[23–29]

In this context, another of our research groups has recently introduced the novel computer-aided molecular design scheme TOMOCOMD-CARDD (acronym of

topological molecular computer design-computer aided 'rational' drug design).[30]

This useful approach based on chemical graph and algebraic theory has been widely used in several QSAR/QSPR studies related to physicochemical and biological properties of chemicals and drugs,[31–35] including studies of nucleic acid–drug interactions[36,37] and discovery of novel antimalarial compounds.[38]

Although some SAR and QSAR studies in tyrosinase inhibitors have been carried out using congeneric datasets,[14,16,39–41] these do not provide lead compounds for future drug development. This kind of data can be only applied to structural lead optimization improving the activity of a selected organic chemical by rearrangement of its substituents. Therefore, a database of heterogeneous compounds may be a successful tool in QSAR research of tyrosinase inhibitors and the discovery of novel lead compounds with different structural features and more effective activity.[42–44]

In the present report, is shown the use of a new set of molecular descriptors (MDs) namely *non-stochastic and stochastic bond-based linear indices*, to find various statistical linear discriminant analysis (LDA) models to discriminate tyrosinase inhibitor compounds from inactive ones. Later, a ligand-based virtual screening of series of compounds was carried out. Finally the in silico selection, isolation, and later pharmacological test of a new set of chemicals are presented, to show the potentialities of the new MDs for drug-discovery processes.

2. Results and discussion

2.1. TOMOCOMD approach

The theory of the bond-based linear indices used in this study was discussed in detail in previous research papers,[45] for an exhaustive overview, see also the Supplementary data.

Molecular fingerprints were generated by means of the interactive program for molecular design and bioinformatic research TOMOCOMD.[30] It is composed of four subprograms; each one of them allows both drawing the structures (drawing mode) and calculating molecular 2D/3D descriptors (calculation mode). The modules are named CARDD (Computer-Aided 'Rational' Drug Design), CAMPS (Computer-Aided Modeling in Protein Science), CANAR (Computer-Aided Nucleic Acid Research), and CABPD (Computer-Aided Bio-Polymers Docking).The CARDD module was selected for drawing all structures and for the computation of non-stochastic and stochastic bond-based linear indices. The main steps for the application of this method in QSAR/QSPR and for drug design can be briefly summarized as follows.

1. Drawing of the molecular pseudographs for each molecule in the dataset, using the drawing mode.
2. Use appropriate weights in order to differentiate the molecular atoms. The weights used in this work are those previously proposed for the calculation of the DRAGON descriptors,[46–48] that is, atomic mass (*M*), atomic polarizability (*P*), atomic Mulliken electronegativity (*K*), van der Waals atomic volume (*V*), plus the atomic electronegativity in Pauling scale (*G*).[49] The values of these atomic labels are shown in Table 1.[46–49]
3. Computation of the total and local (atom and atom-type) bond linear indices of the molecular pseudograph's atom adjacency matrix can be carried out in the software calculation mode, where one can select the atomic properties and the descriptor family before calculating the molecular indices. This software generates a table in which the rows correspond to the compounds, and the columns correspond to the bond-based (both total and local) linear maps or other MD family implemented in this program.

Table 1. Values of the atom weights used for linear indices calculation[71,77–79]

ID	Atomic mass (g/mol)	VdW[a] volume (Å^3)	Mulliken electronegativity	Polarizability (Å^3)	Pauling electronegativity
H	1.01	6.709	2.592	0.667	2.2
B	10.81	17.875	2.275	3.030	2.04
C	12.01	22.449	2.746	1.760	2.55
N	14.01	15.599	3.194	1.100	3.04
O	16.00	11.494	3.654	0.802	3.44
F	19.00	9.203	4.000	0.557	3.98
Al	26.98	36.511	1.714	6.800	1.61
Si	28.09	31.976	2.138	5.380	1.9
P	30.97	26.522	2.515	3.630	2.19
S	32.07	24.429	2.957	2.900	2.58
Cl	35.45	23.228	3.475	2.180	3.16
Fe	55.85	41.052	2.000	8.400	1.83
Co	58.93	35.041	2.000	7.500	1.88
Ni	58.69	17.157	2.000	6.800	1.91
Cu	63.55	11.494	2.033	6.100	1.9
Zn	65.39	38.351	2.223	7.100	1.65
Br	79.90	31.059	3.219	3.050	2.96
Sn	118.71	45.830	2.298	7.700	1.96
I	126.90	38.792	2.778	5.350	2.66

[a] van der Waals.

4. Development of a QSPR/QSAR equation by using several multivariate analytical techniques, for instance, linear discrimination analysis. Therefore, one can find a quantitative relationship between an activity **A** and the bond-based linear fingerprints having, for example, the following appearance:

$$\mathbf{A} = a_0 \boldsymbol{f}_0(w) + a_1 \boldsymbol{f}_1(w) + a_2 \boldsymbol{f}_2(w) + \ldots + a_k \boldsymbol{f}_k(w) + c \quad (1)$$

where **A** is the measured activity, $\boldsymbol{f}_k(w)$ are the kth non-stochastic total bond-based linear indices, and the $\boldsymbol{a}'_k s$ and c are the coefficients obtained by the linear regression analysis.

5. Test of the robustness and predictive power of the QSPR/QSAR equation by using internal [leave-*one*-out (LOO)] and external (using a test set and an external predicting set) validation techniques.

The bond-based TOMOCOMD-CARDD descriptors computed in this study were the following:

(1) kth ($k = 15$) total non-stochastic bond-based linear indices not considering and considering H-atoms in the molecular graph (G) [$\boldsymbol{f}_k(w)$ and $\boldsymbol{f}^{\mathrm{H}}_k(w)$, respectively].
(2) kth ($k = 15$) total stochastic bond-based linear indices not considering and considering H-atoms in the molecular graph (G) [$^{\mathrm{s}}\boldsymbol{f}_k(w)$ and $^{s}\boldsymbol{f}^{\mathrm{H}}_k(w)$, respectively].
(3) k^{th} ($k = 15$) bond-type local (group = heteroatoms: S, N, O) non-stochastic linear indices not considering and considering H-atoms in the molecular graph (G) [$\boldsymbol{f}_{k\mathrm{L}}(w_E)$ and $\boldsymbol{f}^{\mathrm{H}}_{k\mathrm{L}}(w_E)$, correspondingly]. These local descriptors are putative molecular charge, dipole moment, and H-bonding acceptors.
(4) kth ($k = 15$) bond-type local (group = heteroatoms: S, N, O) stochastic linear indices not considering and considering H-atoms in the molecular graph (G) [$^{\mathrm{s}}\boldsymbol{f}_{k\mathrm{L}}(w_E)$ and $^{s}\boldsymbol{f}^{\mathrm{H}}_{k\mathrm{L}}(w_E)$, correspondingly]. These local descriptors are putative molecular charge, dipole moment, and H-bonding acceptors.

2.2. Chemical database selection

In order to assure an adequate extrapolation power in a classification (or any QSAR) model, it is important to take into account one of the most critical aspects, a dataset with a great molecular diversity for it. In our case we have selected 658 compounds for making up the database, having a great degree of structural variability, 246 of them with tyrosinase inhibitor activity considering different inhibition modes and diverse structural patterns, and the rest inactive ones.

The subset of the active chemicals (246 tyrosinase inhibitors) was chosen from the literature, with different structural subsystems, to warrant enough structural diversity; it includes many representative families such as: chalcones,[13] phenolic compounds,[50] kojic acid tripeptides,[51] novel N-substituted *N*-nitrosohydroxylamines,[52] vitamin B_6 compounds,[53] steroids,[54] and so on. A representative sample of the active chemicals selected for this study is depicted in Figure 1. The names of compounds in the database, together with their experimental values taken from the literature, are summarized in Supplementary data (Table 1). The molecular structures of tyrosinase inhibitors are given as Table 2 of Supplementary data. The great degree of variability in the structure of active compounds in the training and prediction sets, as well as their different mechanisms of inhibition, increases the possibilities in the process of selection/design. Therefore, novel leads can be discovered, including compounds with new modes of inhibition of the enzyme activity. The data of tyrosinase inhibitors presented here can be considered as a helpful tool, not only for the theoretical research, but also for the general scientific work in the tyrosinase inhibitor field.

In the case of the inactive compounds selected for both, training and test sets, we chose at random 412 drugs, having a series of other clinical uses. In this inactive group, they were included, for example, antibiotics, antivirals, sedatives/hypnotics, diuretics, anticonvulsants, hemostatics, oral hypoglycemics, antihypertensives, anthelminthics, anticancer, antifungal, and so on, to warrant enough structural diversity.

These reported organic-chemicals were taken from the Negwer Handbook[55] where their names, synonyms and structural formulas can be found. The classification of these compounds as 'inactive' (non-inhibitors of tyrosinase) does not assure that any inhibitory activity does not exist for those organic-chemicals that have not been detected. This problem can be reflected in the results of classification for the series of inactive chemicals.[56]

Besides, the developed LDA-based QSAR models can help us to identify new tyrosinase inhibitors from a combinatorial library or large database of chemicals. This aspect is also shown in this work through a 'simulated' virtual screening.

However, it is necessary to split the active and inactive series into the training and test sets, to find the classification functions. In the first instance, the structural diversity of these datasets should be proved; for that reason, we performed hierarchical cluster analyses (CAs) of the active and inactive series, respectively.[57,58] The statistical software package STATISTICA[59] was used to develop these CAs. The resulting dendrograms are depicted in Figures 2 and 3, using the Euclidean distance (X-axis) and the complete linkage (Y-axis), illustrating the results of the k-NNCA (k-nearest neighbors cluster analysis) developed into active and inactive sets, correspondingly. As it can be observed in both binary trees there are a great number of different structural patterns, which demonstrate the molecular variability of the selected chemicals in this database.

This procedure allows choosing the organic-chemicals for the training and test sets, using a 'rational' way. Because of the difficulty in evaluating the output dendrogram, other kind of CA is usually performed. In this case, we perform two partitional (non-hierarchical)

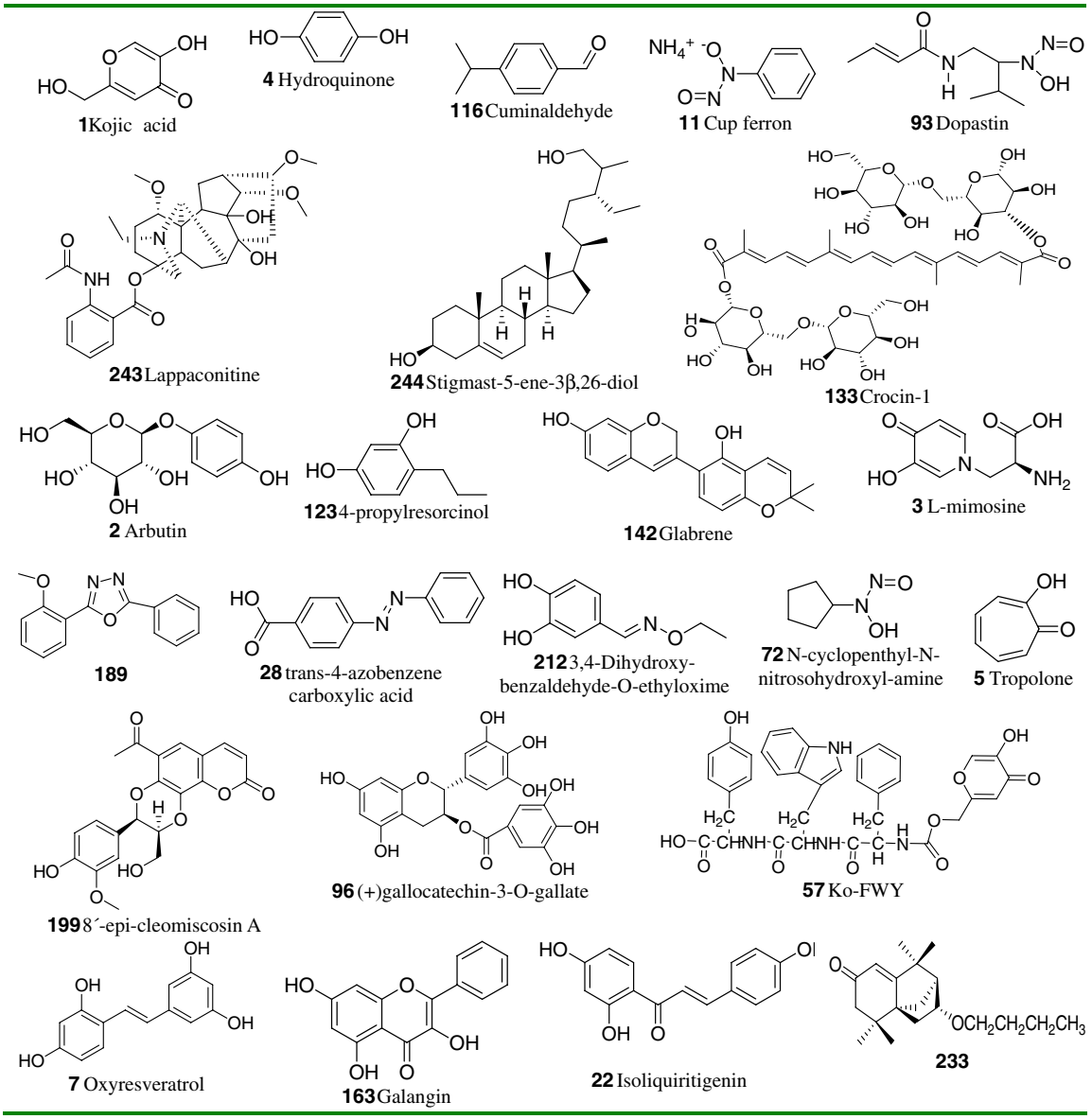

Figure 1. Random, but not exhaustive, sample of the molecular families of tyrosinase inhibitors studied here.

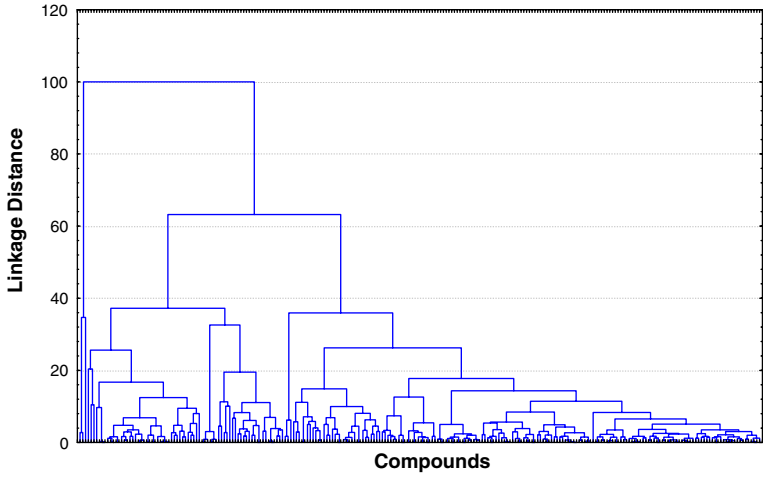

Figure 2. A dendrogram illustrating the results of the hierarchical *k*-NNCA of the set of tyrosinase inhibitors used in the training and prediction sets of the present work.

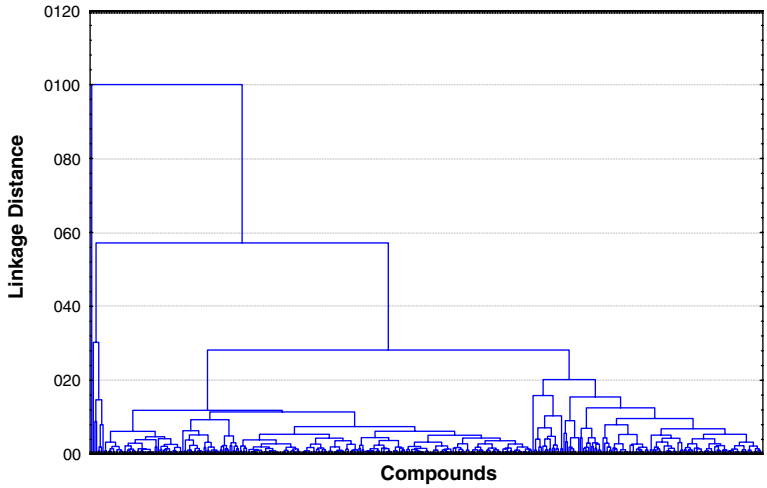

Figure 3. A dendrogram illustrating the results of the hierarchical *k*-NNCA of the set of inactive compounds (non-inhibitors of tyrosinase) used in the training and prediction sets of the present work.

CAs formerly called *k*-MCAs (*k*-means cluster analyses): this procedure permits dividing the whole group into two datasets (training and predicting ones). The main idea of this procedure consists in making a partition of either active or inactive series of chemicals into several statistically representative classes of compounds. This procedure ensures that any chemical subsystems (as determined by the clusters derived from *k*-MCA) will be represented in both compounds' series. This 'rational' design of the training and predicting series permitted us to design both sets that are representative of the whole 'experimental universe.'

A *k*-MCA was made first with active compounds and, afterwards, with inactive ones. The first *k*-MCA (*k*-MCA I) partitioned the tyrosinase inhibitors into 10 clusters. A second *k*-MCA (*k*-MCA II) was realized to split the data of inactive compounds, resulting in 12 clusters; k^{th} non-stochastic bond-based linear indices were used, with all variables showing *p*-levels <0.005 for the Fisher test. The results are depicted in Table 2.

By this way, the selection of the training and prediction sets was performed by taking, in a random way, compounds belonging to each cluster. As we have remarked these 658 chemicals were divided into training set with 478 chosen at random, being 183 of them actives and 295 inactive ones. The remaining subsets composed of 180 chemicals, 63 tyrosinase inhibitors and 117 compounds with different pharmacological uses, were prepared as the test set for the external validation of the classification models. The compounds in the external set were not used in the development of the LDA models. In Figure 4 the above-described whole procedure performed to select a representative sample for the training and test sets through independent CAs is shown graphically.

Table 2. Main results of the *k*-MCAs, for tyrosinase inhibitors and inactive drug-like compounds

Variables	Between SS[a]	Within SS[b]	Fisher ratio (*F*)	*p*-level[c]
Analysis of variance				
Tyrosinase inhibitors clusters (*k*-MCA I)				
${}^{M}f_{0L}(x_E)$	599.37	40.08	392.17	0.00
${}^{P}f^{H}_{0L}(x_E)$	276.96	27.69	262.29	0.00
${}^{M}f_{1L}(x_E)$	613.51	141.56	113.64	0.00
${}^{M}f_{2L}(x_E)$	640.17	82.52	203.42	0.00
${}^{M}f_{4}(x)$	149.61	41.54	94.45	0.00
${}^{K}f_{1L}(x_E)$	233.38	29.07	210.51	0.00
Inactives clusters (*k*-MCA II)				
${}^{M}f_{0L}(x_E)$	2.97	0.90	119.55	0.00
${}^{P}f^{H}_{0L}(x_E)$	519.05	63.84	295.64	0.00
${}^{M}f_{1L}(x_E)$	10.39	1.20	314.43	0.00
${}^{M}f_{2L}(x_E)$	12.20	1.08	412.07	0.00
${}^{M}f_{4}(x)$	626.02	56.32	404.17	0.00
${}^{K}f_{1L}(x_E)$	468.22	47.13	361.26	0.00

[a] Variability between groups.
[b] Variability within groups.
[c] Level of significance. The 0.00 values mean lesser than 0.005.

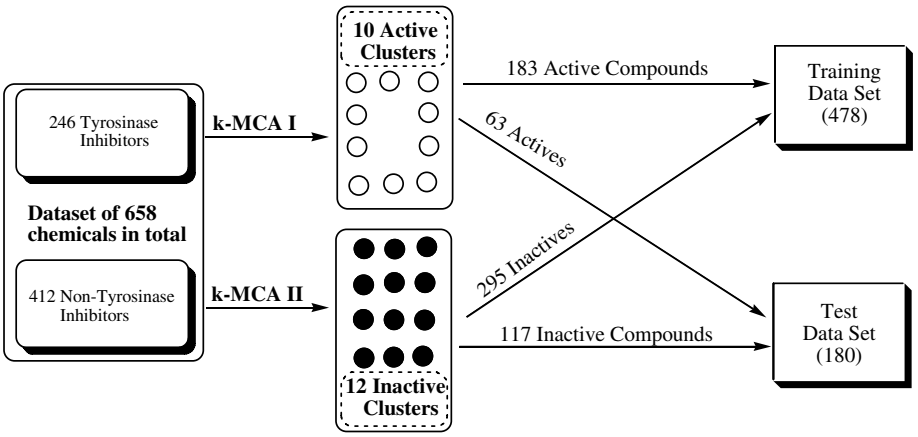

Figure 4. General algorithm used to design training and test sets throughout *k*-MCA.

2.3. Developing discriminant models

The next step, once we perform a representative selection of training series, is to use different statistical techniques to fit discriminant functions. These classification functions permit the classification of chemicals as either active (tyrosinase inhibitors) or inactive, using the linear discriminant analysis (LDA),[60] a method broadly used in drug design.[27,28,35,61–64,56,60,65,66] The total and bond-type non-stochastic and stochastic linear indices were used as independent variables. These new MDs were computed using the weighting schemes previously proposed in Section 2.1: 'TOMOCOMD Approach' (see Table 1).

The classification models are shown in Table 3. In total 12 LDA-based QSAR models were obtained, the first six models using the non-stochastic total and local bond-based linear indices (Eqs. 2–7) and the six remaining models with the stochastic molecular descriptors (Eqs. 8–13). In adittion, in Table 4 we give the prediction performances and the Wilks' statistics (λ), the square of the Mahalanobis distances (D^2), and the Fisher ratio (F) for the LDA-based QSAR models with the training set. The equations showed to be statistically significant at p-level ($p < 0.0001$).

The first five LDA models in both sets were obtained by using each one of the five atomic properties employed as atomic weight (atomic labels) and a sixth model combining molecular indices computed with the whole proposed weighting schemes.

As it can be observed in Table 4, the best results were obtained with the fitted models by using a combination of the weighted schemes, including the non-stochastic and stochastic bond-based linear indices (Eqs. 7 and (13, respectively). These best models correctly classified the 98.95% and 89.75% (accuracy) of the training set. High Matthews' correlation coefficients ($\boldsymbol{C}$ of 0.98 and of 0.78, respectively) are also observed.[67]

In the same Table 4, we also depict most of the parameters commonly used in medical statistics [sensitivity, specificity, and false-positive rate (also known as 'false-alarm rate')] for the whole set of developed models. While the sensitivity is the probability of correctly predicting a positive case, the specificity (also known as 'hit rate') is the probability that a positive prediction is correct.[67]

The best model (Eq. 7) shows a percentage of false actives in this dataset of 0.3%, that is, only 1 inactive compound was classified as actives out of 295 cases. In the group of 183 actives, 4 compounds were misclassified as inactive ones (2.19% misclassification). Although the model of Eq. 13 exhibits good results, no compound for the inactive group was classified as active out of 295. In addition, 1.09% misclassification, 2 compounds were observed as false inactives of a total of 183 actives.

We also checked the linear discriminant canonical statistics: canonical correlation coefficient ($R_{\text{canonical}}$), Chi-square and its p-level [$p(\chi^2)$]; the results are shown in Table 4.[68]

The canonical transformation of the LDA results with non-stochastic (Eq. 7) and stochastic (Eq. 13) bond-based linear fingerprints gives rise to canonical roots with good canonical correlation coefficients of 0.85 and 0.72. The chi-square test permits us to assess the statistical significance of this analysis as having a p-level <0.0001.

2.4. External validation and orthogonalization

Although the statistical parameters in the training dataset had a good behavior, this does not assure the predictive power of the models. The validation process using an external set is one of the most important criteria to evaluate the predictive ability of a QSAR model.[69,70] In this case, the discriminant functions were used to predict the activity of the compounds in the test set. The best two TOMOCOMD-CARDD models (Eqs. 7 and 13) show globally good classifications of 98.89% and 89.44%, respectively, in the prediction series (Table 5). Furthermore, a high value of $\boldsymbol{C}$ can be observed in the Eqs. 7 and 13. Figures 5 and 6 give a plot of the ΔP% (see Experimental Section) for the classification of all compounds in both training and test sets from models **7** and **13**, correspondingly.

Table 3. Discriminant models obtained with total and local non-stochastic and stochastic bond-based linear indices used in this study

LDA-based QSAR models obtained using non-stochastic linear indices

$$\textbf{\textit{Class}} = -1.614 + 1.016 \times 10^{-4}\ {}^{M}f_4(w) - 2.097 \times 10^{-2}\ {}^{M}f^{H}_{0L}(w_E) - 1.687 \times 10^{-2}\ {}^{M}f^{H}_{1L}(w_E) + 4.477 \times 10^{-3}\ {}^{M}f_{0L}(w_E) + 5.351 \times 10^{-3M} f_{1L}(w_E) - 1.311 \times 10^{-3}\ {}^{M}f_{2L}(w_E) \quad (2)$$

$$\text{Class} = -1.481 - 4.814 \times 10^{-2V} f^{H}_0(w) + 0.110^{V} f^{H}_1(w) - 4.316 \times 10^{-2V} f^{H}_2(w) + 7.366 \times 10^{-4V} f^{H}_4(w) + 1.122 \times 10^{-2}\ {}^{V}f_2(w) - 3.120 \times 10^{-4}\ {}^{V}f_4(w) + 2.546 \times 10^{-3}\ {}^{V}f^{H}_{2L}(w_E) + 3.631 \times 10^{-2}\ {}^{V}f_{0L}(w_E) - 2.964 \times 10^{-2}\ {}^{V}f_{1L}(w_E) \quad (3)$$

$$\text{Class} = -0.884 + 1.243^{P} f^{H}_1(w) - 0.670^{P} f^{H}_2(w) + 7.675 \times 10^{-2}\ {}^{P}f^{H}_3(w) - 0.338^{P} f_0(w) + 0.184^{P} f_2(w) - 2.762 \times 10^{-2}\ {}^{P}f_3(w) - 0.474^{P} f^{H}_{0L}(w_E) + 0.162^{P} f^{H}_{1L}(w_E) + 0.746^{P} f_{0L}(w_E) - 0.481^{P} f_{1L}(w_E) + 3.066 \times 10^{-2P} f_{2L}(w_E) \quad (4)$$

$$\textbf{\textit{Class}} = -2.145 + 0.153^{K} f^{H}_1(w) - 5.552 \times 10^{-2}\ {}^{K}f^{H}_2(w) + 3.072 \times 10^{-9}\ {}^{K}f^{H}_{12}(w) + 0.165^{K} f_0(w) + 0.167^{K} f^{H}_{1L}(w_E) - 4.461 \times 10^{-2}\ {}^{K}f^{H}_{2L}(w_E) + 0.113^{K} f_{0L}(w_E) - 0.230^{K} f_{1L}(w_E) + 4.319 \times 10^{-2}\ {}^{K}f_{2L}(w_E) - 1.085 \times 10^{-7}\ {}^{K}f_{10L}(w_E) \quad (5)$$

$$\text{Class} = -2.143 + 0.165^{G} f^{H}_1(w) - 5.998 \times 10^{-2G} f^{H}_2(w) + 3.311 \times 10^{-9}\ {}^{G}f^{H}_{12}(w) + 0.177^{G} f_0(w) + 0.173^{G} f^{H}_{1L}(w_E) - 4.659 \times 10^{-2G} f^{H}_{2L}(w_E) + 0.118^{G} f_{0L}(w_E) - 0.244^{G} f_{1L}(w_E) + 4.608 \times 10^{-2G} f_{2L}(w_E) - 1.171 \times 10^{-7}\ {}^{G}f_{10L}(w_E) \quad (6)$$

$$\text{Class} = -1.317 + 9.556 \times 10^{-5M} f_4(w) + 4.854 \times 10^{-3M} f_{0L}(w_E) + 6.346 \times 10^{-3M} f_{1L}(w_E) - 1.591 \times 10^{-3M} f_{2L}(w_E) - 0.335^{P} f^{H}_{0L}(w_E) - 7.217 \times 10^{-2}\ {}^{K}f_{1L}(w_E) \quad (7)$$

LDA-based QSAR models obtained using stochastic linear indices

$$\text{Class} = -0.957 + 8.321 \times 10^{-2}\ {}^{M}f^{H}_1(w) + 0.169^{M} f^{H}_2(w) - 0.880^{M} f^{H}_4(w) + 1.503^{M} f^{H}_5(w) - 0.906^{M} f^{H}_6(w) + 7.455 \times 10^{-2}\ {}^{M}f^{H}_{15}(w) - 2.912 \times 10^{-2}\ {}^{M}f_1(w) + 3.139 \times 10^{-2}\ {}^{M}f^{H}_{1L}(w_E) + 5.731 \times 10^{-2}\ {}^{M}f_{1L}(w_E) - 7.704 \times 10^{-2}\ {}^{M}f_{2L}(w_E) + 7.140 \times 10^{-2M} f_{14L}(w_E) \quad (8)$$

$$\text{Class} = -0.526 + 0.387^{V} f^{H}_3(w) - 0.485^{V} f^{H}_4(w) + 0.116^{V} f^{H}_{15}(w) - 6.560 \times 10^{-2V} f_0(w) + 5.855 \times 10^{-2V} f_{15}(w) - 5.054 \times 10^{-2}\ {}^{V}f^{H}_{0L}(w_E) + 8.017 \times 10^{-2}\ {}^{V}f^{H}_{1L}(w_E) + 0.216^{V} f^{H}_{8L}(w_E) - 0.258^{V} f^{H}_{15L}(w_E) + 0.102^{V} f_{0L}(w_E) - 0.120^{V} f_{2L}(w_E) \quad (9)$$

$$\text{Class} = -1.197 + 0.349^{P} f^{H}_0(w) + 1.252^{P} f^{H}_1(w) - 7.146^{P} f^{H}_3(w) + 28.133^{P} f^{H}_5(w) - 23.662^{P} f^{H}_6(w) + 1.568^{P} f^{H}_{15}(w) - 1.903^{P} f_1(w) + 2.464^{P} f_4(w) - 2.384^{P} f_9(w) + 1.480^{P} f_{15}(w) - 0.777^{P} f\ {}^{H}_{0L}(w_E) + 0.470^{P} f\ {}^{H}_{5L}(w_E) \quad (10)$$

$$\text{Class} = -1.038 - 0.383^{K} f_9(w) + 0.424^{K} f_{15}(w) + 0.338^{K} f^{H}_{1L}(w_E) + 2.597^{K} f^{H}_{3L}(w_E) - 3.117^{K} f^{H}_{4L}(w_E) + 0.348^{K} f^{H}_{15L}(w_E) + 0.179^{K} f_{0L}(w_E) + 0.318^{K} f_{1L}(w_E) - 2.216^{K} f_{3L}(w_E) + 1.387^{K} f_{4L}(w_E) \quad (11)$$

$$\text{Class} = -1.031 - 0.405^{G} f_9(w) + 0.449^{G} f_{15}(w) + 0.356^{G} f^{H}_{1L}(w_E) + 2.755^{G} f^{H}_{3L}(w_E) - 3.306^{G} f^{H}_{4L}(w_E) + 0.373^{G} f^{H}_{15L}(x_E) + 0.190^{G} f_{0L}(x_E) + 0.342^{G} f_{1L}(x_E) - 2.381^{G} f_{3L}(x_E) + 1.494^{G} f_{4L}(x_E) \quad (12)$$

$$\text{Class} = -0.275 - 4.813 \times 10^{-2M} f_{1L}(w_E) - 0.356^{P} f^{H}_0(w) + 0.297^{P} f^{H}_{15}(w) - 0.445^{K} f^{H}_{1L}(w_E) + 0.271^{K} f_{0L}(w_E) - 0.353^{K} f_{3L}(w_E) + 1.716^{G} f^{H}_{3L}(w_E) - 1.981^{G} f^{H}_{4L}(w_E) \quad (13)$$

Table 4. Prediction performances and statistical parameters for LDA-based QSAR models in the training set

Models[a]	Matthews correlation coefficient (C)	Accuracy 'Q_{Total}' (%)	Specificity (%)	Sensitivity 'hit rate' (%)	False positive rate (%)	Wilks' λ	D^2	F	Chi-square (χ^2)	Canonical$R$ (R_{can})[b]
LDA-based QSAR models obtained using non-stochastic linear indices										
Eq. 2 (6)	0.96	98.11	99.4	95.6	0.3	0.29	10.08	187.8	577.7	0.84
Eq. 3 (9)	0.82	91.63	88.6	89.6	7.1	0.48	4.58	56.5	346.6	0.72
Eq. 4 (11)	0.82	91.63	91.3	86.3	5.1	0.48	4.51	45.3	342.3	0.72
Eq. 5 (10)	0.80	90.79	89.3	86.3	6.4	0.47	4.85	53.7	349.1	0.72
Eq. 6 (10)	0.80	90.79	89.3	86.3	6.4	0.47	4.85	53.7	347.1	0.72
Eq. 7 (6)	**0.98**	**98.95**	**99.4**	**97.8**	**0.3**	**0.28**	**10.97**	**204.3**	**606.2**	**0.85**
LDA-based QSAR models obtained using stochastic linear indices										
Eq. 8 (11)	0.78	89.75	86.4	86.9	8.5	0.48	4.57	45.9	345.4	0.72
Eq. 9 (11)	0.72	86.40	80.7	84.7	12.5	0.5	4.27	42.9	329.0	0.71
Eq. 10 (12)	0.68	85.15	83.7	76.0	9.2	0.53	3.72	34.2	297.4	0.68
Eq. 11 (10)	0.73	87.24	82.8	84.2	10.8	0.49	4.47	49.5	340.6	0.72
Eq. 12 (10)	0.73	87.03	82.4	84.2	11.2	0.49	4.47	49.5	340.3	0.72
Eq. 13 (8)	**0.78**	**89.75**	**86.0**	**87.4**	**8.8**	**0.48**	**4.63**	**64.4**	**349.9**	**0.72**

[a] Between brackets the quantity of variables of the models.
[b] Canonical correlation coefficient obtained from the linear discriminant canonical analysis.

Table 5. Prediction performances for LDA-based QSAR models in the test set

Models	Matthews Corr. Coefficient (C)	Accuracy 'Q_{Total}' (%)	Specificity (%)	Sensitivity 'hit rate' (%)	False positive Rate (%)
LDA-based QSAR models obtained using non-stochastic linear indices					
Eq. 2	0.96	98.33	100	95.2	0
Eq. 3	0.74	88.33	85.0	81.0	7.7
Eq. 4	0.78	90.00	86.9	84.1	6.8
Eq. 5	0.72	87.22	83.3	79.4	8.5
Eq. 6	0.72	87.22	83.3	79.4	8.5
Eq. 7	**0.98**	**98.89**	**100**	**96.8**	**0**
LDA-based QSAR models obtained using stochastic linear indices					
Eq. 8	0.73	87.22	79.4	85.7	12.0
Eq. 9	0.70	84.44	71.1	93.7	20.5
Eq. 10	0.69	86.11	82.8	76.2	8.6
Eq. 11	0.72	87.22	80.3	84.1	11.1
Eq. 12	0.71	86.67	79.1	84.1	12.0
Eq. 13	**0.77**	**89.44**	**82.4**	**88.9**	**10.3**

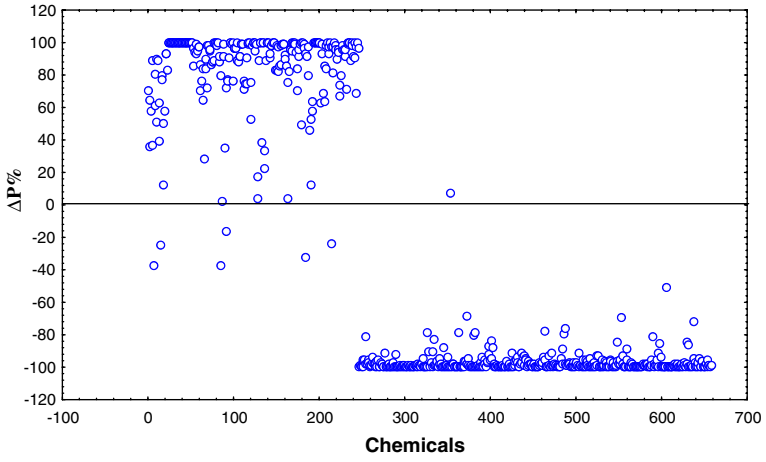

Figure 5. Plot of the ΔP% from Eq. 7 (using non-stochastic linear indices) for each compound in the training and test sets. Compounds 1–183 and 184–246 are active (tyrosinase inhibitors) in training and test sets, respectively; chemicals 247–541 and 542–658 are inactive (non-inhibitors of tyrosinase) in both training and test sets, correspondingly.

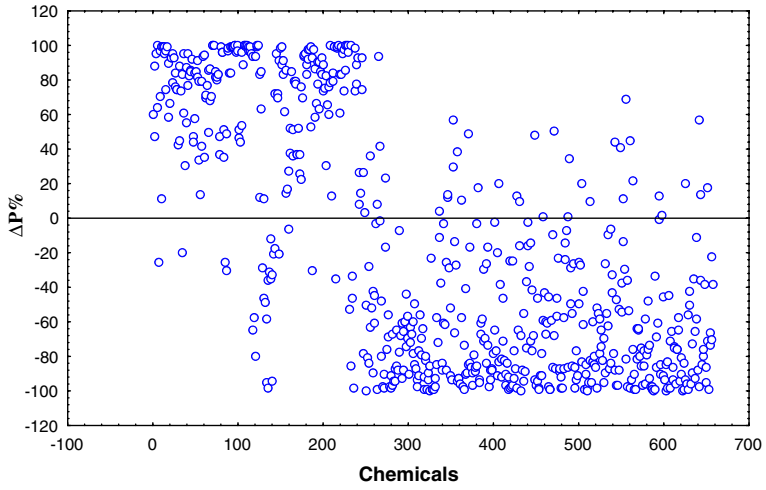

Figure 6. Plot of the $\Delta P\%$ from Eq. 13 (using stochastic linear indices) for each compound in the training and test sets. Compounds 1–183 and 184–246 are active (tyrosinase inhibitors) in training and test sets, respectively; chemicals 247–541 and 542–658 are inactive (non-inhibitors of tyrosinase) in both training and test sets, correspondingly.

In Table 5 are summarized the results of the statistical parameters using the non-stochastic and stochastic bond-based linear indices. Consequently, the models validated by these results can be used to generate a virtual in silico screening.

The classification results (including the canonical scores) for all the compounds (active and inactive ones) in the training database with all the models are depicted in Tables 3–6 of Supplementary data. In the same way, the results of classification using all the developed models for the test set are shown in Tables 7–10 (Supplementary data). The SMILES notation for these active and inactive dataset is given in Tables 11 and 12 of Supplementary data, respectively.

However, when we analyze the molecular descriptors included in the best LDA-based QSAR models, a strong interrelation between these molecular fingerprints is observed (data not shown). To overcome this difficulty, in the interpretation of the QSAR model, we used the Randić's orthogonalization process of the molecular descriptors. This process is an approach in which molecular descriptors are transformed in such a way that they do not mutually correlate, to avoid the exclusion of descriptors included in the model.[64,71–76] As expected, the non-orthogonal descriptors and derived orthogonal descriptors contain the same information, as well as the statistical parameters continue being the same for both models. Moreover, the coefficients of the QSAR model based on orthogonal descriptors allow us to interpret the correlation coefficient and to evaluate the role of individual fingerprints in the QSAR model.

The results of the orthogonalization process for the non-stochastic and stochastic bond-based linear indices are shown in Eqs. 14 and 15, respectively.

$$\begin{aligned}\text{Class} = & -2.771 + 3.467^{1}O(^{\mathrm{M}}\boldsymbol{f}_{0\mathrm{L}}(w_{\mathrm{E}})) \\ & - 2.558^{2}O(^{\mathrm{P}}\boldsymbol{f}^{\mathrm{H}}_{0\mathrm{L}}(w_{\mathrm{E}})) \\ & + 2.344^{3}O(^{\mathrm{M}}\boldsymbol{f}_{1\mathrm{L}}(w_{\mathrm{E}})) \\ & - 4.584^{4}O(^{\mathrm{M}}\boldsymbol{f}_{2\mathrm{L}}(w_{\mathrm{E}})) + 1.328^{5}O(^{\mathrm{M}}\boldsymbol{f}_{4}(w)) \\ & - 2.267^{6}O(^{\mathrm{K}}\boldsymbol{f}_{1L}(w_{E})) \\ & N = 478,\ \lambda = 0.28,\ D^{2} = 10.97, \\ & F = 204.3,\ \text{canonical } R = 0.85, \\ & \chi^{2} = 606.2,\ Q_{\mathrm{Total}} = 98.95\%,\ C = 0.98. \end{aligned} \quad (14)$$

$$\begin{aligned}\text{Class} = & \ 0.019 - 0.858^{1}O(^{\mathrm{M}}\boldsymbol{f}_{1\mathrm{L}}(w_{\mathrm{E}})) \\ & + 4.643^{2}O(^{K}\boldsymbol{f}^{\mathrm{H}}_{1\mathrm{L}}(w_{\mathrm{E}})) \\ & - 5.081^{3}O(^{\mathrm{G}}\boldsymbol{f}^{\mathrm{H}}_{4\mathrm{L}}(w_{\mathrm{E}})) \\ & + 0.492^{4}O(^{\mathrm{P}}\boldsymbol{f}^{\mathrm{H}}_{15}(w)) - 3.062^{5}O(^{\mathrm{P}}\boldsymbol{f}^{\mathrm{H}}_{0}(w)) \\ & + 15.568^{6}O(^{\mathrm{G}}\boldsymbol{f}^{\mathrm{H}}_{3\mathrm{L}}(w_{\mathrm{E}})) \\ & + 1.578^{7}O(^{\mathrm{K}}\boldsymbol{f}_{0\mathrm{L}}(x_{\mathrm{E}})) \\ & - 3.965^{8}O(^{\mathrm{K}}\boldsymbol{f}_{3\mathrm{L}}(x_{\mathrm{E}})) \\ & N = 478,\ \lambda = 0.48,\ D^{2} = 4.63, \\ & F = 64.4,\ \text{canonical } R = 0.72, \\ & \chi^{2} = 349.9,\ Q_{\mathrm{Total}} = 89.75\%,\ C = 0.78. \end{aligned} \quad (15)$$

Here, we used the symbols $^{m}O[\boldsymbol{f}_{k}(w)]$, where the superscript m expresses the order of importance of the variable $\boldsymbol{f}_{k}(w)$ after a preliminary forward-stepwise analysis and O means orthogonal.

It is important to stand out that, as expected, the orthogonal descriptor-based models coincide with the collinear (i.e., ordinary) TOMOCOMD-CARDD descriptors-based models in all the statistical parameters. Again, the statistical coefficients of LDA-QSARs

λ, F, C and accuracy are the same whether we use a set of non-orthogonal descriptors or the corresponding set of orthogonal indices. Notice that in the process of orthogonalization the data were standardized so that each variable has a mean of zero and a standard deviation of 1, because the different molecular descriptors used entirely 'different types of scales.'

2.5. New tyrosinase inhibitors through ligand-based virtual screening

Recent advances in the pharmaceutical industry are focused on resolving the ancient problem of massive cost in the development of new drugs. These novel approaches to answer the major challenges most commonly expected in drug discovery are based on automation and information technologies. The mentioned tools have provided the pharmaceutical industry with platforms to translate clinical liabilities into simple, fast, and cost-effective in vitro screening assays, applicable to the early phases of drug discovery.[77]

High-throughput screening (HTS) has emerged as the paradigm chosen by the pharmaceutical companies as a solution to query large databases of compounds searching for a desired activity.[78] Despite the great advances in the biological screening of large number of organic-chemicals by using the HTS techniques, the process of drug discovery is still an arduous task.

Therefore, the development of computational methods that can serve for this purpose is necessary. Computers scientists may regard virtual high-throughput screening (vHTS) as a highly sophisticated way, and as a strongly simplified simulation of their high-throughput screening assays. Indeed, vHTS attempts to integrate computer science with biophysics using their synergy: the flexibility, cost-effectiveness, and speed of computational algorithms and biophysical knowledge of molecular recognition.[19]

Therefore, they have emerged the computational methods permitting theoretical-in silico-evaluations of such activity for virtual libraries of chemicals before these compounds are synthesized in the laboratory, as an interesting alternative to the HTS in the bottleneck of the drug-discovery pipeline.

The term 'virtual' encompasses in the pharmaceutical industry all aspects related to the electronic business paradigm. This type of 'virtual' strategy developed in the last five years has the potential to increase the efficiency of drug commercialization. Nevertheless, it should not be confused with 'in silico' drug discovery approaches enlarging the use of computational software to generate and evaluate molecules that are also 'virtual.'[79]

This in silico search is presented as an alternative to the screening approaches to drug discovery. The main feature that any theoretical approach should have is its ability to identify new active compounds from databases of chemicals. Therefore, the computational methods can permit the 'virtual essay' of tyrosinase inhibitory activity from virtual libraries of chemicals, before the biological tests are carried out, or even these compounds are synthesized in the laboratory. The process mentioned is known today as computational (virtual or in silico) screening.[17–28]

With the aim of evaluating the applicability of the QSAR models obtained in the present work, we carried out a simulated virtual screening of tyrosinase inhibitors. A dataset of 85 compounds, whose names are given in Table 6, was evaluated in the ligand-based virtual screening. The collected organic-chemicals are reported as active in the literature (see the last column of Table 6: Ref.). The molecular structures of these compounds are depicted in Table 13 of Supplementary data.

First, we develop a k-NNCA to observe the molecular diversity in the database of the virtual screening. In the dendrogram of Figure 7 a great number of different subsets can be visualized proving the large molecular variability of the chemicals.

The results of the classification of the compounds in the screening are depicted in Table 6. Together with these, we show the results of $\Delta P\%$ (posterior classification probabilities) and canonical scores of the compounds using all the models developed in the present report (see Table 14 of Supplementary data). In addition, pictorial representations of the good classification of these compounds obtained with the best two models **7** and **13** are given in Figures 8 and 9.

As it can be seen in Figure 8, a total of 36 compounds [57.64% (49/85)] were misclassified by Eq. 7. However, the Eq. 13, had a adequate behavior in this external series, showing a 92.94% of good classification (79/85). Other obtained models showed intermediate results according to its accuracy in training and test sets. The values of predictions were checked out from recent reports in the literature (see the last column of Table 6: **Ref.**). Therefore, in the virtual screening experiments using this computational system is very important to use the following criterion to choose a compound as good-inhibitor candidate (hit or lead): (1) If the 75% of the total 12 LDA-based QSAR models pick the compound as active, then we classify the compound with the biological activity under consideration.

By this way, the following step would be the inclusion of these 'novel' compounds in the training set and then to carry out new discrimination models. This novel discrimination functions can have some variability from the previous one, due to the inclusion of a new structural pattern, but they should be able to create a broad spectrum of compounds that allow recognizing a greater number of structural subsets from databases as tyrosinase inhibitors. Hence, the improvement of the quality of the models incorporating new compounds is considered as an iterative process of great impact in the construction of adequate datasets.

Table 6. Results of the virtual screening

Compound[a]	Class[b]	Ref.[c]
Active compounds (tyrosinase inhibitors)		
1 *p*-Nitrophenol	++++++	A
	++++++	B
2 3-(3,4-Dihydroxyphenyl)-L-alanine	++++++	C
	++++++	
3 3-Amino-4-hydroxybenzoic acid	++++++	C
	++++++	
4 4-Amino-3-hydroxybenzoic acid	++++++	C
	++++++	
5 3,4-Diaminobenzoic acid	++++++	C
	++++++	
6 3-Aminobenzoic acid	**++++++**	C
	++++++	
7 4-Aminobenzoic acid	++++++	C
	++++++	
8 4,6-*O*-Hexahydroxy diphenylglucose	++++++	D
	++++++	
9 Tunicamycin	+−−−−+	E
	++−+++	
10 Methyl *p*-coumarate	++++++	F
	++++++	
11 *o*-Phenylphenol	−++++−	F
	++++++	
12 Phenylhydroquinone	++++++	F
	++++++	
13 Chamaecin	++++++	F
	++−+++	G
14 Stearyl glycyrrhetinate	++++++	H
	++++++	
15 2-(4-Methylphenyl)- 1,3-	−−−−−−	I
selenazol-4-one	+−−−−−	J
16	−−−−−−	I
	+−−−−−	
17	−−−−−−	I
	+−+−−−	
18	−−−−−−	I
	+−−−−−	
19 3-Fluorotyrosine	++++++	K
	++++++	
20 *N*-Acetyltyrosine	+−−−−+	K
	++++++	
21 *N*-Formyltyrosine	++++++	K
	++++++	
22 Gentisic acid	++++++	L
	++++++	
23 6-BH_4	+−−−−+	M
	−+−−−+	
24 7-BH_4	+−−−−+	M
	−+−−−+	
25 Propylparaben	++++++	N
	++−+++	
26 Phenylalanine	++++++	K
	++++++	
27 Dithiothreitol	++++++	O
	++++++	
28 Azelaic acid	+−−+++	P
	++−+++	
29 Undecandioic acid	+−−+++	P
	++−+++	
30 Suberic acid	+−−+++	P
	++−+++	
31 Sebacic acid	+−−+++	P
	++−+++	
32 Dodecandioic acid	+−−+++	P
	++−+++	
33 Tridecandioic acid	+−−+++	P
	++−+++	

Table 6 (*continued*)

Compound[a]	Class[b]	Ref.[c]
34 Traumatic acid	+−−+++	P
	++−+++	
35 Pantothenic acid	++++++	K
	++−+++	
36 5-(Hydroxymethyl)-2-furfural	+−−−−+	Q
	+−−+++	R
37 Hinokitiol	−++++−	S
	++++++	
38 Penicillamine	+++−−+	T
	++++++	
39 Toluic acid	++++++	A
	++++++	
40	++++++	U
	++++++	
41	++++++	U
	++++++	
42 3,5-Dihydroxy-	++++++	V
4′-*O*-Methoxystilbene	++++++	
43 *p*-Hydroxybenzoic acid	++++++	W
	++++++	
44 *o*-Hydroxybenzoic acid	++++++	W
	++++++	
45 Cysteine	++++++	X
	++++++	
46 Methimazole	+−+−−+	X
	+++−−−	
47 BMY-28438	−++++−	X
	++++++	
48 Captopril	+−+−−+	Y
	+−−+++	
49 Yohimbine	+++−−+	Z
	++−+++	
50 4-(Phenylazo)phenol	++++++	a
	++++++	
51 SACat	++++++	A
	++++++	
52 NPACat	++++++	A
	++++++	
53 DNPACat	++++++	a
	++++++	
54 EDTA	+−−−−+	b
	−−−+++	
55 Dodecyl gallate	++++++	c
	++−+++	
56 Gallic acid	++++++	c
	++++++	
57 (±)-Flavanone	−−−+−−−	d
	+++−−−	
58 (−)-Pinocembrin	−++++−	d
	++++++	
59 (±)-Naringenin	−++++−	d
	++++++	
60 (+)-Dihydromorin	−++++−	d
	++++++	
61 Flavone	−++++−	d
	++++−+	
62 Myricetin	−++++−	d
	++++++	
63 Artocarpin	−++++−	d
	++++++	
64 Artocarpesin	−++++−	d
	++++++	
65 Isoartocarpesin	−++++−	d
	++++++	
66 (−)-Angolensin	−++++−	d
	++++++	

Table 6 (*continued*)

Compound[a]	Class[b]	Ref.[c]
67 Pinosylvin	−++++− ++++++	d
68 4-Prenyloxyresveratrol	−++++− ++++++	d
69 26	−++++− ++++++	d
70 27	−++++− ++++++	D
71 28	−++++− ++++++	D
72 29	−++++− ++−+++	D
73 30	−++++− ++++++	D
74 31	−++++− ++++++	D
75 32	−++++− ++++++	D
76 34	−++++− ++++++	D
77 35	−++++− ++++++	D
78 36	−++++− ++++++	D
79 37	−++++− ++++++	d
80 38	−++++− ++++++	d
81 39	−++++− ++−+++	d
82 40	−++++− ++++++	d
83 41	−++++− ++−+++	d
84 2′-*O*-Feruloylaloesin	−++++− ++++++	e
85 Barbaloin	−++++− ++++++	e

[a] The molecular structures of these tyrosinase inhibitors are given as Supplementary data (see Table 13).

[b] Results of the classification of compounds in this set: (i) Above, classification of each compound using the obtained models with non-stochastic bond-based linear indices in the following order: Eqs. 2–7; and (ii) Below; classification of each compound using the obtained models with stochastic bond-based linear indices in the following order Eqs. 8–13.

[c] References taken from the literature: [A]Bubacco, L.; van Gastel, M.; Groenen, E. J. J.; Vijgenboom, E.; Canters, G. W. *J. Biol. Chem.* **2003**, *278*, 7381–7389. [B]van Gastela, M.; Bubaccob, L.; Groenena, E. J. J.; Vijgenboomc, E.; Cantersc, G. W. *FEBS Lett.* **2000**, *474*, 228–232. [C]Gasowskaa, B.; Kafarskia, P.; Wojtasek, H. *Biochim. Biophys. Acta* **2004**, *1673*, 170–177. [D]http://open.cacb.org.tw/index.php (2005-03-03 09:09:51). [E]Takahashi, H.; Parsons, P. G. *J. Invest. Dermatol.* **1992**, *98*, 481–487. [F]Kubo, I.; Niheia, K.; Tsujimoto, K. *Bioorg. Med. Chem.* **2004**, *12*, 5349–5354. [G]Nihei, K-I.; Yamagiwa, Y.; Kamikawab, T.; Kubo, I. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 681–683. [H]Um, S.-J.; Park, M.-S.; Park, S.-H.; Han, H.-S.; Kwonb, Y.-J.; Sin, H.-S. *Bioorg. Med. Chem.* **2003**, *11*, 5345–5352. [I]Barlocco, D.; Barrett, D.; Edwards, P.; Langston, S.; Pérez-Pérez, M. J.; Walker, M.; Weidner, J.; Westwell, A. *Drug Discovery Today* **2003**, *8*, 372–373. [J]Koketsu, M.; Choi, S. Y.; Ishihara, H.; Lim B. O.; Kim, H.; Kim, S. Y. *Chem. Pharm. Bull.* (*Tokyo*) **2002**, *12*, 1594–1596. http://www.thecosmeticsite.com/formulating/959621.htlm (April-00). [L]Curto, E. V.; Kwong, C.; Hermersdorfer, H.; Glatt, H.; Santis, C.; Virador, V.; Hearing, V. J.; Dooley, T. P. *Biochem. Pharmacol.* **1999**, *57*, 663–672. [M]Wood, J. M.; Schallreuter-Wood, K. U.; Lindsey, N. J.; Callaghan, S.; Gardner, M. L. G. *Biochem. Biophys. Res. Commun.* **1995**, *206*, 480–485. [N]Hori, I.; Nihei, K.-I.; Kubo, I. *Phytother. Res.* **2004**, *18*, 475–479. [O]Naish-Byfield, S.; Cooksey, C. J.; Riley, P. A. *Biochem. J.* **1994**, *304*, 155–162. [P]Nazzaro-Porro, M.; Passi, S. *J. Invest. Dermatol.* **1978**, *71*, 205–208. [Q]Sharma, V. K.; Choi, J.; Sharma, N.; Choi, M.; Seo, S.-Y. *Phytother. Res.* **2004**, *18*, 841–844. [R]Kang, H. S.; Choi, J. H.; Cho, W. K.; Park, J. C.; Choi, J. S. *Arch. Pharm. Res.* **2004**, *7*, 742–750. [S]Sakuma, K.; Ogawa, M.; Sugibayashi, K.; Yamada, K.; Yamamoto, K. *Arch. Pharm. Res.* **1999**, *4*, 335–339. [T]Lovstad, R. A. *Biochem. Pharmacol.* **1976**, *25*, 533–535. [U]Kubo, I.; Kinst-Hori, I.; Yokokawa, Y. *J. Nat. Prod.* **1994**, *57*, 545–551. [V]Regev-Shoshani, G.; Shoseyov, O.; Bilkis, I.; Kerem, Z. *Biochem. J.* **2003**, *374*, 157–163. [W]Bernard, P.; Berthon, J-Y. *Int. J. Cosmetic Sci.* **2000**, *22*, 219–226. [X]Imada, C.; Sugimoto, Y.; Makimura, T.; Kobayashi, T.; Hamada, N.; Watanabe, E. *Fish Sci.* **2001,** *67*, 1151–1156. [Y]Espín, J. C.; Wichers, H. J. *Biochim. Biophys. Acta*. **2001**, *1544*, 289-300. [Z]Fuller, B. B.; Drake, M. A.; Spaulding, D. T.; Chaudry, F. *J. Invest. Dermatol.* **2000**, *114*, 268–276. [a]Borojerdi, S. S.; Haghbeen, K.; Karkhane, A. A.; Fazli, M.; Sabouryc, A. A. *Biochem. Biophys. Res. Commun.* **2004**, *314*, 925–930. [b]Kong, K.-H.; Hong, M.-P.; Choi, S.-S.; Kim, Y.-T.; Cho, S.-H. *Biotechnol. Appl. Biochem.* **2000**, *31*, 113–118. [c]Kubo, I.; Chen, Q.-X.; Nihei, K.-I. *Food Chem.* **2003**, *81*, 241–247. [d]Shimizu, K.; Kondo, R.; Sakai, K. *Planta Med.* **2000**, *66*, 11–15. [e]Yagi, A.; Kanbara, T.; Morinobu, N. *Planta Med.* **1987**, 515–517.

Several drug-like compounds were identified by the QSAR models as possible tyrosinase inhibitors. There is great variability in the functions of these chemicals, for example, one antirheumatic compound (Penicillamine, also used as copper chelating agent in Wilson's hepatolenticular degeneration disease), one antihyperthyroid chemical (Methimazole), one antihypertensive agent (Captopril), one mydriatic lead (Yohimbine; also used as pharmacological probe for the study of alpha-2 adrenoreceptor as well as in the treatment of impotence), one antibacterial molecular entity (BMY-28438), one analgesic and antiinflammatory compound (Gentisic acid), and so on. A great diversity is observed in the molecular structures of these chemicals. Among them, one can find from well-known drugs with different biological properties to natural products.

The above-shown result is one of the most important validations for the models developed here, because we have demonstrated that the present algorithm is able to detect as active a series of compounds from a library, and these chemicals have shown the predicted activity. It is important to recall the fact that most of these drugs selected from the ligand-based virtual screening can have well-established methods of synthesis, as well as their toxicological, pharmacodynamical, and pharmaceutical properties are well known.

2.6. In silico discovery of new tyrosinase inhibitors and experimental assays

The main importance and usefulness of QSAR models developed here is the selection of lead compounds from a large group of chemical-organics. In order to exemplify the possibilities of the TOMOCOMD-CARDD

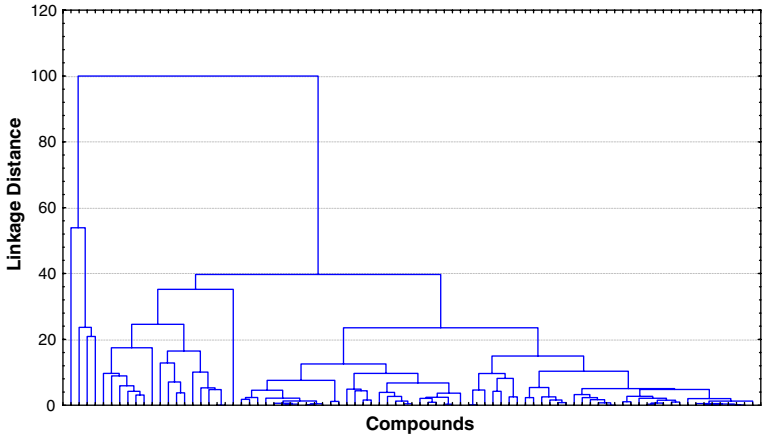

Figure 7. A dendrogram illustrating the results of the hierarchical *k*-NNCA of the set of active chemicals used for evaluating the predictive ability of the QSAR models for ligand-based virtual screening.

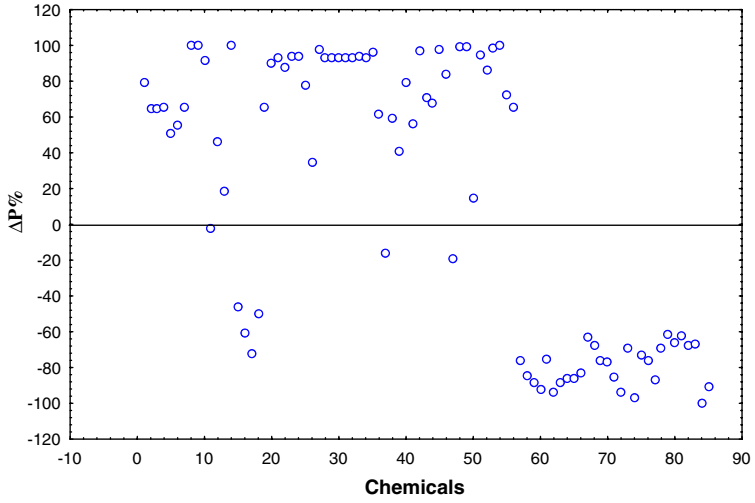

Figure 8. Plot of the $\Delta P\%$ from Eq. 7 (using non-stochastic linear indices) for each compound selected in virtual screening protocols.

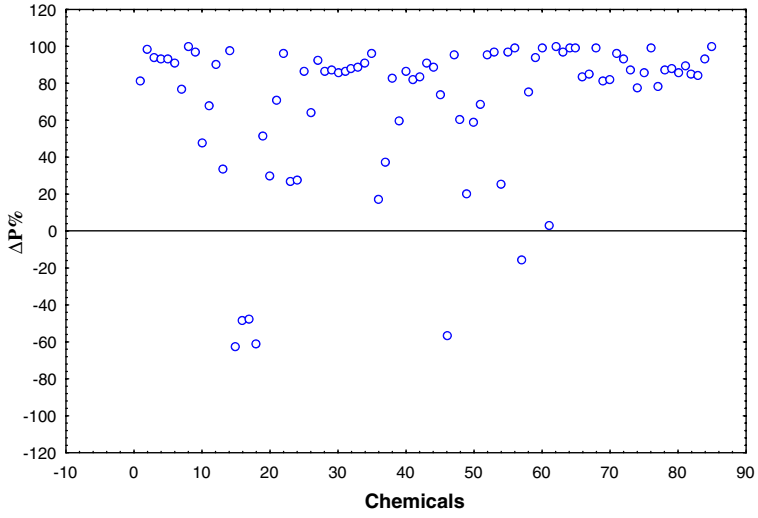

Figure 9. Plot of the $\Delta P\%$ from Eq. 13 (using stochastic linear indices) for each compound selected in virtual screening protocols.

approach for the in silico screening, a search for new bioactive chemicals was carried out.

As it was mentioned in other part of this report, one of our research teams has focused mainly on trial–error searching for new tyrosinase inhibitors.[14–16] Together with this fact, we are also identifying new drug candidates using computational screening (based on QSAR techniques). In order to find promising active agents, we selected a pool of compounds not yet described in the literature as tyrosinase inhibitors. In the following step, we perform in silico essays from a library of ethylsteroids isolated and characterized from natural sources, looking for activity against the tyrosinase enzyme, using the QSAR models obtained with the TOMOCOMD-CARDD method.

We evaluated six compounds with the LDA-based QSAR models and, in order to corroborate the predictions, the chemicals were isolated through simple methods, and in vitro assays against the enzyme were made. The $\Delta P\%$ values and canonical scores of the compounds in the data using all the discriminant functions are depicted in Table 7.

A good agreement is observed between the experimental antityrosinase activity and theoretical predictions for all the compounds. In the study of the inhibitory activity all six compounds showed effectiveness in mushroom tyrosinase inhibition (see Table 7).[80] Two compounds, **ES3** (IC_{50} = 7.89 μM) and **ES4** (IC_{50} = 5.95 μM), exhibited lesser activity than standard tyrosinase inhibitor L-mimosine, but more power than Kojic acid (IC_{50} = 16.67 μM), another reference drug. The remaining chemicals, **ES1** (IC_{50} = 2.61 μM), **ES2** (IC_{50} = 1.53 μM), **ES5** (IC_{50} = 3.46 μM) and **ES6** (IC_{50} = 1.72 μM), exhibited higher activities when compared with L-mimosine (IC_{50} = 3.68 μM). The structures of the compounds are shown in Figure 10.

It must be highlighted here that all compounds have an $IC_{50} < 10$ μM, the minor value to consider a compound as a *hit* for drug discovery. For that reason, we also made a *k*-NNCA for all the active compounds of the training, test, virtual screening, and the new bioactive chemicals. The hierarchical cluster analysis was developed to compare similarities among the novel discovered compounds and the complete active database. The dendrogram of Figure 11 shows the great diversity of subsystems in the set.

An exhaustive analysis of each cluster reveals that new bioactive agents were included in the same cluster that an ethylsteroid family, **234–238** and **244–246**, compounds belonging in the main to the training set. This result can be considered adequate because the novel chemicals are ethylsteroids too. In adition, it could be observed from the cluster that **ES2** (IC_{50} = 1.53 μM) is at the same distance as that compound **238** (IC_{50} = 1.25 μM), reasonable, taking into account that they share similar structural shapes and strong tyrosinase inhibitory activity (see Figure 2 of Supplementary data and Fig. 10).

From these results, we can conclude that the models developed here permit, at least, the identification of these nuclei bases that are simple derivatives of known class of tyrosinase inhibitors as novel tyrosinase inhibitors which may be used to prevent or treat pigmentation disorders.

All this new small library of compounds may be used as starting point for further optimization and refinement of novel compounds with potent tyrosinase activity. **ES2** can be selected in search for drug-like compounds with such activity, after examining the pharmacological, toxicity, pharmacokinetic properties, and good activity in clinical animal essays. Finally, it is important to remark that our aim in this study is to show how the models can be used for potential drug discovery.

Table 7. Results of ligand-based in silico screening and tyrosinase inhibitory activities of new ethylsteroid compounds

Compound*	$\Delta P\%^a$	Scores[a]	$\Delta P\%^b$	Scores[b]	$\Delta P\%^c$	Scores[c]	$\Delta P\%^d$	Scores[d]	$\Delta P\%^e$	Scores[e]	$\Delta P\%^f$	Scores[f]	IC_{50} ± SEM[g] (μM)
ES1	**65,84**	**−1,02**	**96,90**	**2,42**	**97,43**	**−2,52**	**95,67**	**−2,13**	**95,56**	**−2,13**	**66,11**	**−1,01**	2.61 ± 0.0373
	84,84	*1,31*	*96,74*	−2,21	*77,70*	*1,55*	*83,62*	*1,37*	*83,05*	*1,35*	*82,65*	−1,23	
ES2	**69,18**	**−1,06**	**97,37**	**2,49**	**96,82**	**−2,42**	**97,87**	**−2,46**	**97,81**	**−2,47**	**67,87**	**−1,03**	1.53 ± 0.0011
	69,62	*0,94*	*88,43*	−1,57	*62,22*	*1,23*	*83,50*	*1,37*	*82,89*	*1,35*	*87,43*	−1,40	
ES3	**63,18**	**−0,99**	**97,10**	**2,45**	**97,80**	**−2,59**	**94,22**	**−1,95**	**94,09**	**−1,94**	**64,58**	**−1,00**	7.89 ± 0.0013
	88,75	*1,46*	*97,14*	−2,27	*92,27*	*2,14*	*83,73*	*1,38*	*83,16*	*1,36*	*79,74*	−1,15	
ES4	**58,49**	**−0,94**	**97,49**	**2,51**	**97,81**	**−2,59**	**98,02**	**−2,47**	**97,99**	**−2,48**	**59,60**	**−0,95**	5.95 ± 0.0008
	90,41	*1,54*	*98,11*	−2,47	*79,79*	*1,61*	*93,51*	*1,83*	*93,31*	*1,82*	*94,85*	−1,83	
ES5	**66,73**	**−1,03**	**97,54**	**2,52**	**97,28**	**−2,49**	**97,15**	**−2,28**	**97,07**	**−2,28**	**66,40**	**−1,02**	3.46 ± 0.0105
	76,74	*1,09*	*89,46*	−1,62	*85,71*	*1,80*	*83,60*	*1,37*	*82,99*	*1,35*	*85,15*	−1,31	
ES6	**62,40**	**−0,98**	**97,86**	**2,59**	**97,28**	**−2,49**	**99,05**	**−2,81**	**99,03**	**−2,82**	**61,63**	**−0,97**	1.72 ± 0.0009
	80,98	*1,19*	*93,32*	−1,85	*66,66*	*1,31*	*93,46*	*1,83*	*93,25*	*1,82*	*96,33*	−1,99	

[a,b,c,d,e,f] $\Delta P\% = [P(\text{Active}) - P(\text{Inactive})] \times 100$ as well as canonical scores of each compound in this set: (i) Above in bold, classification of each compound using the obtained models with non-stochastic linear indices in the following order: Eqs. 2–7; and (ii) Below in italic; classification of each compound using the obtained models with stochastic linear indices in the following order Eqs. 8–13. [g] IC_{50} are the 50% inhibitory concentrations against the enzyme tyrosinase and SEM is the standard error of the mean.

* The molecular structures of these chemicals are shown in Figure 10.

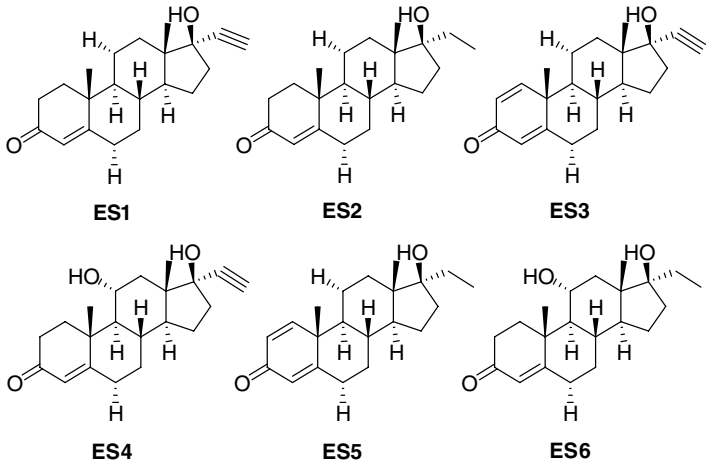

Figure 10. Molecular structure of the new ethylsteroid compounds.

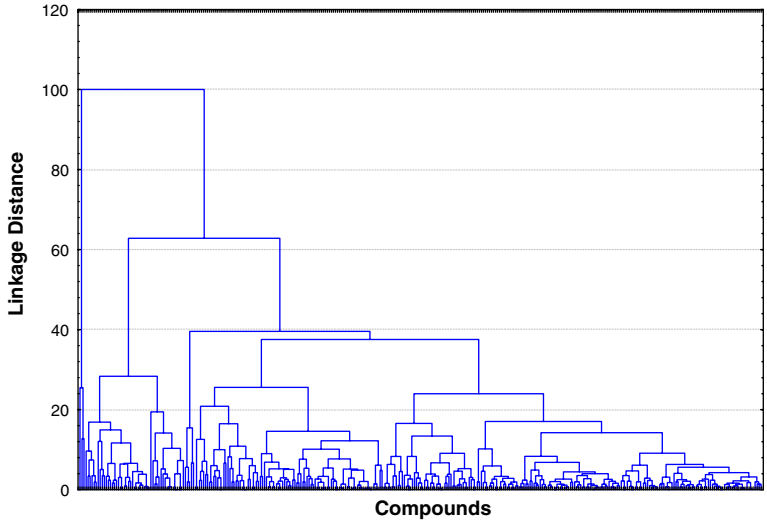

Figure 11. A dendrogram illustrating the results of the hierarchical *k*-NNCA of the set of all active chemicals (tyrosinase inhibitors) included in training, test, virtual screening, and new active ethylsteroids discovered in the present work.

2.7. Bond-based linear indices versus other simple molecular fingerprints

Besides, a comparison with other standard methods was also carried out to corroborate the performance of our own molecular descriptors. With this purpose were used three standard descriptor families, the molecular walk counts (2D), functional groups (1D) and the atom centered fragments (1D).[81] The LDA-based QSAR models obtained for these three molecular fingerprints are shown in Table 8.

The statistical parameters for these three models and for our best two non-stochastic and stochastic bond-based descriptors are shown in Tables 9 and 10, for the training and prediction series, respectively. As can be observed the Eqs. 16–18 present low values of accuracy, specificity, and sensitivity than those presented by our models **7** and **13**. In this sense the reliability of these simple molecular fingerprints used here [molecular walk counts (2D), functional groups (1D), atom centered fragments (1D)] cannot provide an adequate discrimination between the tyrosinase inhibitors and inactive ones. Therefore, the use of most complex molecular descriptors like the bond-based linear indices to describe the biological activity is necessary.

3. Concluding remarks

The research involving the discovery of new tyrosinase inhibitors is considered an impacting field in pharmaceutical, cosmetic, food, and more recently agrochemical industries. This fact is due to their potential applications in such areas as food additives, skin depigmentation agents, insect pests, melanogenesis disorders, and so on.[9–13,41] The broad spectrum of the tyrosinase enzyme makes it a useful target for drug discovery of new inhibitors with the bioactivity.

Methods saving in effectiveness and rationality have become a principal objective for the pharmaceutical re-

Table 8. Discriminant models obtained with the 0D–2D dragon descriptors

$$\text{Class} = -0.830 - 2.899 \times 10^{-2}\text{SRW05} + 2.094 \times 10^{-2}\text{MWC04} \quad (16)$$

$$\begin{aligned}\text{Class} = {} & -0.260 + 2.623n\text{CO} - 1.104n\text{NR2} - 1.762n\text{SO2}N - 0.181n\text{Cq} + 0.772n\text{COOR} \\ & + 1.062n\text{CrHR} - 2.244n\text{NO2} + 0.691\text{nRORPh} - 7.790 \times 10^{-2}n\text{CrR2} + 1.103\text{n} \\ & = \text{CR2} + 0.698n\text{C} = N - 0.256n\text{CONR2}\end{aligned} \quad (17)$$

$$\begin{aligned}\text{Class} = {} & -0.1085 - 0.1520\text{S} - 110 - 1.575\text{S} - 107 - 1.211\text{S} - 108 - 0.655\text{Br} - 094 \\ & - 0.694\text{Cl} - 090 - 1.082\text{P} - 117\end{aligned} \quad (18)$$

Table 9. Prediction performances and statistical parameters for LDA-based QSAR models in the training set

Models[a]	Matthews correlation coefficient (C)	Accuracy 'Q_{Total}' (%)	Specificity (%)	Sensitivity 'hit rate' (%)	False positive rate (%)	Wilks' λ	D^2	F	Chi-Square (χ^2)	Canonical R
LDA-based QSAR models obtained using the dragon descriptors										
Eq. 16 (2)	0.072	54.64	42.25	49.45	42.12	0.96	0.19	10.4	20.4	0.21
Eq. 17 (12)	0.54	78.48	86.36	52.19	5.13	0.65	2.25	20.6	200.0	0.59
Eq. 18 (6)	0.35	58.01	47.68	96.15	65.75	0.88	0.59	11.0	61.9	0.35
Eq. 7 (6)	**0.98**	**98.95**	**99.4**	**97.8**	**0.3**	**0.28**	**10.97**	**204.3**	**606.2**	**0.85**
Eq. 13 (8)	**0.78**	**89.75**	**86.0**	**87.4**	**8.8**	**0.48**	**4.63**	**64.4**	**349.9**	**0.72**

Table 10. Prediction performances for LDA-based QSAR models in the test set

Models[a]	Matthews correlation coefficient (C)	Accuracy 'Q_{Total}' (%)	Specificity (%)	Sensitivity 'hit rate' (%)	False positive rate (%)
LDA-based QSAR models obtained using the dragon descriptors					
Eq. 16	0.25	65.36	50.7	55.6	29.3
Eq. 17	0.48	77.09	73.9	54.0	10.3
Eq. 18	0.41	61.45	47.7	96.8	57.8
Eq. 7	**0.98**	**98.89**	**100**	**96.8**	**0**
Eq. 13	**0.77**	**89.44**	**82.4**	**88.9**	**10.3**

search. In this sense, the use of in silico approaches has emerged as a replacement alternative to in vivo test assays. For example, the virtual high-throughput-screening techniques could do parallel testing of libraries of compounds, accelerating the lead generation process or even new drug candidates.[82,83]

In this way, other approaches have been proposed to evaluate and complement the HTS in virtual assays. The introduction and use of the described graph theoretical MDs are also attractive and efficient for research in drug design. Taken all these facts and, in spite of some criticism, the vTHS together with the classification-based QSAR models can become a common tool for testing compounds before entering more refined and costly assays to be conducted later, thus speeding up drug discovery.

In this way and knowing that most of the tyrosinase inhibitors discovered until present have been discovered by trial–error methods, we have shown the biological in silico evaluation through QSAR models of new compounds isolated and characterized from herbal plants.

In the current report, the usefulness of the TOMOCOMD-CARDD MDs *non-stochastic and stochastic bond-based linear indices* was shown to discriminate novel active compounds from inactive ones as tyrosinase inhibitors. The obtained discriminate functions were applied to pools of chemicals in the simulated virtual screening of compounds with the activity under study, exhibiting adequate performances. This new method is proposed for increasing the speed of prediction of the biological property considered here, permitting much better results for the in silico discovery of new candidates as possible lead compounds, making use of a minimum of resources. The data collected in this work can be used by all scientists in natural-product, medicinal or theoretical chemistry area of tyrosinase inhibitor researches.

At the same time, the novel MDs were used with the LDA technique as an experimental screening of the novel ethylsteroids. This was proved experimentally, with an in vitro pharmacological essay of the isolated and characterized compounds. In this case, the new six chemicals presented activity on mushroom tyrosinase, which proves that the TOMOCOMD-CARDD approach is a

useful tool in the rational design of novel pharmacologically active compounds.

The present algorithm constitutes a step forward in the search for efficient ways of discovering new tyrosinase inhibitors. This strategy will 'deliver substantial benefits and act as the bedrock for NCE selection'[79] shedding some new light into the drug-discovery pipeline, improving the quality during the *hit*-to-*lead* stage of novel bioactive compounds, looking for more potent-safety-selective tyrosinase inhibitors, which may be used to prevent or treat pigmentation disorders.

4. Experimental

4.1. Chemoinformatic tools

Calculations were carried out on a PC Pentium-4 2.0 GHz. The CARDD module implemented in the TOMOCOMD software was used in the calculation of total and local non-stochastic and stochastic bond-based linear indices for the dataset of chemicals. The atom weights used were the same as those for the calculation of the DRAGON descriptors.[46–48]

The three standard descriptor families molecular walk counts (2D), functional groups (1D) and the atom centered fragments (1D) were calculated using the Dragon Software.[84]

4.2. Chemometric methods

4.2.1. *k*-Means cluster analysis (*k*-MCA). The statistical software package STATISTICA[59] was used to develop the *k*-MCA.[59] The number of members in each cluster and the standard deviation of the variables in the cluster (kept as low as possible) were taken into account, to have an acceptable statistical quality of data partitions into the clusters. The values of the standard deviation (SS) between and within clusters, of the respective Fisher ratio and their p level of significance, were also examined.[57,58] Finally, before carrying out the cluster processes, all the variables were standardized. In standardization, all the values of selected variables (molecular descriptors) were replaced by standardized values, which are computed as follows: Std. score = (raw score − mean)/Std. deviation.

4.2.2. Linear discriminant analysis (LDA). LDA was carried out with the STATISTICA software.[59] The considered tolerance parameter (proportion of variance that is unique to the respective variable) was the default value for minimum acceptable tolerance, which is 0.01. A forward-stepwise search procedure was fixed as the strategy for variable selection. The principle of parsimony (Occam's razor) was taken into account as a strategy for model selection. In connection, we selected the model with the highest statistical significance, but having as few parameters (a_k) as possible. The quality of the models was determined by examining Wilks' λ parameter (U statistic), the square Mahalanobis distance (D^2), the Fisher ratio (F), and the corresponding p-level [$p(F)$] as well as the percentage of good classification in the training and test sets. Models with a proportion between the number of cases and variables in the equation lower than 5 were rejected. The biological activity was codified by a dummy variable 'Class'. This variable indicates the presence of either an active compound (Class = 1) or an inactive compound (Class = −1). The classification of cases was performed by means of the posterior classification probabilities. By using the models, one compound can then be classified as either active, if $\Delta P\% > 0$, being $\Delta P\% = [P(\text{Active}) - P(\text{Inactive})]100$, or inactive, otherwise. P(Active) and P(Inactive) are the probabilities with which the equations classify a compound as either active or inactive, respectively.

The statistical robustness and predictive power of the obtained model were assessed using a prediction (test) set. Finally, the calculation of percentages of global good classification (accuracy), sensibility, specificity (also known as 'hit rate'), false positive rate (also known as 'false alarm rate'), and Matthews' correlation coefficient (MCC) in the training and test sets permitted the assessment of the model.[67]

4.2.3. Orthogonalization of descriptors. In this study, the Randić method of orthogonalization was used.[64,71–76] This orthogonalization process of molecular descriptors was introduced by Randić several years ago as a way to improve the statistical interpretation of the models by using interrelated indices. This method has been described in detail in several publications.

Thus, we will give only a general overview here. As a first step, an appropriate order of orthogonalization was considered following the order with which the variables were selected in the forward-stepwise search procedure of the statistical analysis.[68] The first variable (V_1) is taken as the first orthogonal descriptor ${}^1O(V_1)$, and the second one (V_2) is orthogonalized with respect to it [${}^2O(V_2)$]. The residual of its correlation with ${}^1O(V_1)$ is that part of the descriptor V_2 not reproduced by ${}^1O(V_1)$. Similarly, from the regression of V_3 versus ${}^1O(V_1)$, the residual is the part of V_3 that is not reproduced by ${}^1O(V_1)$, and it is labeled ${}^1O(V_3)$. The orthogonal descriptor ${}^3O(V_3)$ is obtained by repeating this process in order to make it orthogonal to ${}^2O(V_2)$ also. The process is repeated until all variables are completely orthogonalized, and the orthogonal variables are then used to obtain the new model.[64,71–76]

Because the different molecular descriptors included here used entirely 'different types of scales,' the data were standardized so that each variable has a mean of 0 and a standard deviation of 1. In standardization, all the values of selected variables (molecular descriptors) were replaced by standardized values, which are computed as follows: Std. score = (raw score − mean)/Std. deviation.

4.3. Chemical methods

The isolation and characterization of the ethylsteroids, their biological studies, and cross references have been reported by others of our research team.[85]

4.4. In vitro assay of tyrosinase activity

Tyrosinase inhibition assay was performed with kojic acid and L-mimosine as standard inhibitors for the tyrosinase in a 96-well microplate format using a SpectraMax 340 micro-plate reader (Molecular Devices, CA, USA) according to the method developed by Hearing.[80] Briefly, the compounds were first screened for the *o*-diphenolase inhibitory activity of tyrosinase using L-DOPA as substrate. All the active inhibitors from the preliminary screening were subjected to IC_{50} studies. Compounds were dissolved in methanol to a concentration of 2.5%. Thirty units of mushroom tyrosinase (28 nM from Sigma Chemical Co., USA) were first pre-incubated with the test compounds in 50 nM Na-phosphate buffer (pH 6.8) for 10 min at 25 °C. Then the L-DOPA (0.5 mM) was added to the reaction mixture and the enzymatic reaction was monitored by measuring the change in absorbance at 475 nm (at 37 °C) due to the formation of the DOPAchrome for 10 min. The percentage of inhibition of the enzyme was calculated as follows, by using MS Excel®™ 2000 (Microsoft Corp., USA) based program developed for this purpose:

$$\text{Percent inhibition} = [(B - S)/B] \times 100 \qquad (19)$$

Here, B and S are the absorbances for the blank and samples, respectively. After the screening of the compounds, 50 percent inhibitory concentrations (IC_{50}) were also calculated. All the studies have been carried out at least in triplicate, and the results represent the mean ± SEM (standard error of the mean). Kojic acid and L-mimosine were used as standard inhibitors for the tyrosinase and both of them were purchased from Sigma Chem. Co., USA.

Acknowledgments

One of the authors (M.-P. Y) thanks Prof. Dr. Ramón García Domenech for the revision of manuscript. Their numerous comments and suggestions on the manuscript which resulted in a significant improvement of the material. The same author acknowledges the Valencia University for kind hospitality during the second semester of 2006. M.-P. Y thanks are also given to the Generalitat Valenciana, (Spain) for partial financial support as well as the program 'Estades Temporals per an Investigadors Convidats' for a fellowship to work at Valencia University (2006–2007). Some authors' thanks support from Spanish MEC (Project Reference: SAF2006-04698). M.T.H.K is the recipient of a grant from MCBN-UNESCO (grant no. 1056), and fellowships from CIB (Italy) and Associasione Veneta per la Lotta alla Talassemia (AVTL, Italy). F.T. acknowledges financial support from the Spanish MEC DGI (Project No.CTQ2004-07768-C02-01/BQU) and Generalitat Valenciana (DGEUI INF01-051 and INFRA03-047, and OCYT GRUPOS03-173.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2006.10.067.

References and notes

1. Hearing, V. J.; Jimenez, M. *Pigment Cell Res.* **1989**, *2*, 75–85.
2. Curto, E. V.; Kwong, C.; Hermersdorfer, H.; Glatt, H.; Santis, C.; Virador, V., Jr.; Hearing, V. J.; Dooley, T. P. *Biochem. Pharmacol.* **1999**, *57*, 663–672.
3. Sanchez-Ferrer, A.; Rodriguez-Lopez, J. N.; Garcia-Canovas, F.; Garcia-Carmona, F. *Biochim. Biophys. Acta* **1995**, *1247*, 1–11.
4. Frenk, E. In *Melasma: New Approaches to Treatment*; Martin Dunitz: London, 1995; pp. 9–15.
5. Dooley, T. P. In *Drug Discovery Approaches for Developing Cosmeceuticals: Advanced Skin Care and Cosmetic Products*; Hori, W., Ed.; International Business Communications: Southborough, MA, 1997.
6. Dooley, T. P. *J. Dermatol. Treat.* **1997**, *7*, 188–200.
7. Kojima, S.; Yamaguchi, H.; Morita, K.; Ueno, Y. *Biol. Pharm. Bull.* **1995**, *18*, 1076–1080.
8. Verallo-Rowell, V. M.; Verallo, V.; Graupe, K.; Lopez-Villafuerte, L.; Garcia-Lopez, M. . *Acta Derm.-Venereol* **1989**, *143*, 58–61.
9. Takiwake, H.; Shirai, S.; Kohono, H.; Soh, H.; Arase, S. *J. Invest. Dermatol.* **1994**, *103*, 642–646.
10. Kimbrough-Green, C. K. *Arch. Dermatol.* **1994**, *130*, 727–733.
11. Kubo, I.; Kinst-Hori, I.; Kubo, Y.; Yamagiwa, Y.; Kamikawa, T.; Haraguchi, H. *J. Agric. Food Chem.* **2000**, *48*, 1393–1399.
12. Nihei, K.; Yamagiwa, Y.; Kamikawa, T.; Kubo, I. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 681–683.
13. Khatib, S.; Nerya, O.; Musa, R.; Shmuel, M.; Tamir, S.; Vaya, J. *Bioorg. Med. Chem.* **2005**, *13*, 433–441.
14. Khan, M. T.; Choudhary, M. I.; Khan, K. M.; Rani, M.; Atta ur, R. *Bioorg. Med. Chem.* **2005**, *13*, 3385–3395.
15. Choudhary, M. I.; Sultan, S.; Khan, M. T. H.; Yasin, A.; Shaheen, F.; Atta-ur-Rahman *Nat. Prod. Res.* **2004**, *18*, 529–535.
16. Ahmad, V. U.; Ullah, F.; Hussain, J.; Farooq, U.; Zubair, M.; Khan, M. T.; Choudhary, M. I. *Chem. Pharm. Bull. (Tokyo)* **2004**, *52*, 1458–1461.
17. Watson, C. *Biosilico* **2003**, *1*, 83–84.
18. Xu, J.; Hagler, A. *Molecules* **2002**, *7*, 566–700.
19. Seifert, H. J. M.; Wolf, K.; Vitt, D. *Biosilico* **2003**, *1*, 143–149.
20. Kubinyi, H. *J. Braz. Chem. Soc.* **2002**, *13*, 717–726.
21. Dixit, K. S.; Mitra, S. N. *CRIPS* **2002**, *3*, 1–7.
22. Manly, C. J.; Louise-May, S.; Hammer, J. D. *Drug Discovery Today* **2001**, *6*, 1101–1110.
23. Lajiness, M. In *Computaional Chemical Graph Theory*; Rouvray, D. H., Ed.; Nova Science: New York, 1990.
24. Estrada, E.; Peña, A.; Garcia-Domenech, R. *J. Comput. Aided Mol. Des.* **1998**, *12*, 583–595.
25. Estrada, E.; Uriarte, E.; Montero, A.; Teijeira, M.; Santana, L.; De Clercq, E. *J. Med. Chem.* **2000**, *43*, 1975–1985.

26. Gonzales-Diaz, H.; Marrero Ponce, Y.; Hernadez, I.; Bastida, I.; Tenorio, E.; Nasco, O.; Uriarte, E.; Castanedo, N.; Cabrera, M. A.; Aguila, E.; Marrero, O.; Morales, A.; Perez, M. *Chem. Res. Toxicol.* **2003**, *16*, 1318–1327.
27. de Julian-Ortiz, J. V.; de Alapont, C. G.; Ríos-Santamarina, I.; Garcia-Domenech, R.; Galvez, E. *J. Mol. Graphics Modell.* **1998**, *16*, 14–18.
28. Montero-Torres, A.; Vega, M. C.; Marrero-Ponce, Y.; Rolon, M.; Gomez-Barrio, A.; Escario, J. A.; Aran, V. J.; Martinez-Fernandez, A. R.; Meneses-Marcel, A. *Bioorg. Med. Chem.* **2005**, *13*, 6264–6275.
29. Marrero Ponce, Y. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2010–2026.
30. Marrero-Ponce, Y.; Romero, V. TOMOCOMD software. Central University of Las Villas; 2002. TOMOCOMD (topological molecular computer design) for Windows, version 1.0 is a preliminary experimental version; in future a professional version can be obtained upon request to Y. Marrero: yovanimp@qf.uclv.edu.cu or ymarrero77@yahoo.es.
31. Marrero-Ponce, Y. *Molecules* **2003**, *8*, 687–726.
32. Marrero-Ponce, Y.; Huesca-Guillen, A.; Ibarra-Velarde, F. *J. Mol. Struct. (THEOCHEM)* **2005**, *717*, 67–79.
33. Meneses-Marcel, A.; Marrero-Ponce, Y.; Machado-Tugores, Y.; Montero-Torres, A.; Pereira, D. M.; Escario, J. A.; Nogal-Ruiz, J. J.; Ochoa, C.; Aran, V. J.; Martinez-Fernandez, A. R.; Garcia Sanchez, R. N. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 3838–3843.
34. Vega, M. C.; Montero-Torres, A.; Marrero-Ponce, Y.; Rolon, M.; Gomez-Barrio, A.; Escario, J. A.; Aran, V. J.; Nogal, J. J.; Meneses-Marcel, A.; Torrens, F. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1898–1904.
35. Marrero-Ponce, Y.; Medina-Marrero, R.; Martinez, Y.; Torrens, F.; Romero-Zaldivar, V.; Castro, E. A. *J. Mol. Model.* **2006**, *12*, 255–271.
36. Marrero-Ponce, Y.; Nodarse, D.; González, H. D.; Ramos de Armas, R.; Romero-Zaldivar, V.; Torrens, F.; Castro, E. *Int. J. Mol. Sci.* **2004**, *5*, 276–293.
37. Marrero Ponce, Y.; Castillo Garit, J. A.; Nodarse, D. *Bioorg. Med. Chem.* **2005**, *13*, 3397–3404.
38. Marrero-Ponce, Y.; Iyarreta-Veitia, M.; Montero-Torres, A.; Romero-Zaldivar, C.; Brandt, C. A.; Avila, P. E.; Kirchgatter, K.; Machado, Y. *J. Chem. Inf. Model* **2005**, *45*, 1082–1100.
39. Kubo, I.; Chen, Q.; Nihei, K. *Food Chem.* **2003**, *81*, 241–247.
40. Shimizu, K.; Kondo, R.; Sakai, K. *Planta Med.* **2000**, *66*, 11–15.
41. Li, W.; Kubo, I. *Bioorg. Med. Chem.* **2004**, *12*, 701–713.
42. Casañola-Martin, G. M.; Khan, M. T.; Marrero-Ponce, Y.; Ather, A.; Sultankhodzhaev, M. N.; Torrens, F. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 324–330.
43. Marrero-Ponce, Y.; Khan, M. T. H.; Casañola-Martín, G. M.; Ather, A.; Sultankhodzhaev, M. N.; Torrens, F. *QSAR Comb. Sci.*, accepted for publication.
44. Marrero-Ponce, Y.; Khan, M. T. H.; Casañola-Martín, G. M.; Ather, A.; Sultankhodzhaev, M. N.; Torrens, F.; Rotondo, R. *ChemMedChem*, submitted for publication.
45. Marrero-Ponce, Y.; Torrens, F.; Rotondo, R. *(See ECSOC-9, Conference Hall G, G-015* http://www.usc.es/congresos/ecsoc/9/ECSOC9.HTM).
46. Pauling, L. In *The Nature of Chemical Bond*; Cornell University Press: Ithaca, New York, 1939; pp 2–60.
47. Todeschini, R.; Gramatica, P. *Perspect. Drug Disc. Des.* **1998**, *9–11*, 355–380.
48. Consonni, V.; Todeschini, R.; Pavan, M. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 682–692.
49. Kier, L. B.; Hall, L. H. In *Molecular Connectivity in Structure–Activity Analysis*; Research Studies Press: Letchworth, UK, 1986.
50. Khan, K. M.; Maharvi, G. M.; Abbaskhan, A.; Hayat, S.; Khan, M. T. H.; Makhmoor, T.; Choudhary, M. I.; Shaheen, F.; Atta-ur-Rahman *Helv. Chim. Acta* **2003**, *86*, 457–464.
51. Kim, H.; Choi, J.; Cho, J. K.; Kim, S. Y.; Lee, Y. S. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 2843–2846.
52. Shiino, M.; Watanabe, Y.; Umezawa, K. *Bioorg. Chem.* **2003**, *31*, 129–135.
53. Yokochi, N.; Morita, T.; Yagi, T. *J. Agric. Food Chem.* **2003**, *51*, 2733–2736.
54. Şabudak, T.; Khan, M. T. H.; Choudhary, M. I.; Oksuz, S. *Nat. Prod. Res.* **2006** DOI: doi:10.1080/14786410500196821.
55. Negwer, M. In *Organic-Chemical Drugs and their Synonyms*; Akademie-Verlag: Berlin, 1987.
56. Estrada, E.; Peña, A. *Bioorg. Med. Chem.* **2000**, *8*, 2755–2770.
57. Mc Farland, J. W.; Gans, D. J. In *Chemometric Methods in Molecular Design*; Waterbeemd, H., Ed.; VCH: New York, 1995; pp 295–307.
58. Johnson, R. A.; Wichern, D. W. In *Applied Multivariate Statistical Analysis*; Prentice-Hall: New Jersey, 1988.
59. STATISTICA (data analysis software system), v. S. I., 2001. www.statsoft.com.
60. van de Waterbeemd, H. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 265–288.
61. Marrero-Ponce, Y.; Montero-Torres, A.; Zaldivar, C. R.; Veitia, M. I.; Perez, M. M.; Sanchez, R. N. *Bioorg. Med. Chem.* **2005**, *13*, 1293–1304.
62. Marrero-Ponce, Y.; Castillo-Garit, J. A.; Olazabal, E.; Serrano, H. S.; Morales, A.; Castanedo, N.; Ibarra-Velarde, F.; Huesca-Guillen, A.; Sanchez, A. M.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2005**, *13*, 1005–1020.
63. Marrero-Ponce, Y.; Cabrera, M. A.; Romero, V.; González, D. H.; Torrens, F. *J. Pharm. Pharm. Sci.* **2004**, *7*, 186–199.
64. Estrada, E.; Vilar, S.; Uriarte, E.; Gutierrez, Y. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1194–1203.
65. Marrero-Ponce, Y.; Díaz, H. G.; Romero, V.; Torrens, F.; Castro, E. A. *Bioorg. Med. Chem.* **2004**, *12*, 5331–5342.
66. Gálvez, J.; García, R.; Salabert, M. T.; Soler, R. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 520–525.
67. Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A.; Nielsen, H. *Bioinformatics* **2000**, *16*, 412–424.
68. Ford, M.-G.; Salt, D.-W. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995; pp 283–292.
69. Wold, S.; Erikson, L. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: New York, 1995; pp 309–318.
70. Golbraikh, A.; Tropsha, A. *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.
71. Randić, M. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 311–320.
72. Randić, M. *New J. Chem.* **1991**, *15*, 517–525.
73. Randić, M. *J. Mol. Struct. (THEOCHEM)* **1991**, *233*, 45–59.
74. Lučić, B.; Nikolić, S.; Trinajstić, N.; Jurić, D. *J. Chem. Inf. Comput. Sci.* **1995**, 532–538.
75. Klein, D. J.; Randić, M.; Babić, D.; Lučić, B.; Nikolić, S.; Trinajstić, N. *Int. J. Quantum Chem.* **1997**, *63*, 215–222.
76. Estrada, E.; Uriarte, E. *Curr. Med. Chem.* **2001**, *8*, 1573–1588.

77. Whitebread, S.; Hamon, J.; Bojanic, D.; Urban, L. *Drug Discovery Today* **2005**, *10*, 1421–1433.
78. Horrobin, D. F. *J. Roy. Soc. Med.* **2000**, *93*, 341–345.
79. Ekins, S.; Boulanger, B.; Swaan, P. W.; Hupcey, M. A. Z. *J. Comput. Aided Mol. Des.* **2002**, *16*, 381–401.
80. Hearing, V. J.. In *Methods in Enzymology*; Academic: New York, 1987; Vol. 142, p 154.
81. Todeschini, R.; Consonni, V. In *Handbook of Molecular Descriptors*; Wiley-VCH: Germany, 2000.
82. Balls, M.; Van Zeller, A.-M.; Halder, M. E. In *Progress in the Reduction and Refinement and Replacement of Animal Experimentation*; Elselvier: Amsterdam, 2000.
83. Oprea, T. I. *J. Comput. Aided Mol. Des.* **2002**, *16*, 325–334.
84. Todeschini, R.; Consonni, V.; Pavan, M., Dragon Software version 2.1, 2002.
85. Choudhary, M. I.; Sultan, S.; Khan, M. T.; Rahman, A. U. *Steroids* **2005**, *70*, 798–802.