Structure−Activity Relationship Studies of Substituted 17α-Acetoxyprogesterone Hormones

R. S. Braga,* R. Vendrame, and D. S. Galvão

Instituto de Física Gleb Wataghin, UNICAMP, CP 6165, CEP 13083-970, Campinas, SP, Brasil

Received April 14, 2000

Recently a new methodology, called electronic indices methodology (EIM), based on local density of state calculations (LDOS) using topological and semiempirical methods, was proposed to identify the biological activity of polycyclic aromatic hydrocarbons (PAHs). In this work we apply the concepts of the EIM approach to classify the progestational activity of 21 17α-acetoxyprogesterones (steroid hormones) (APs). The EIM approach pointed to a few descriptors, which correctly classify the active/inactive compounds of this class (≈90%). We show that these descriptors arise naturally from principal component analysis (PCA) and neural network (NN) calculations. Moreover, using only the parameters from EIM, instead of a large set of descriptors that have been used before to describe the biological activity of these hormones, we slightly improve and simplify PCA and NN results. Finally, the molecular region related to the chemical activity of these hormones naturally appears in our theoretical analysis, from the local density of states of the frontier orbitals. This shows the generality of the principles of EIM approach, and confirms that the combination of these distinct methodologies can be an efficient and powerful tool in the structure−activity studies of many different classes of compounds.

1. INTRODUCTION

Substituted 17α-acetoxyprogesterones (APs)[1] belong to a class of very important compounds called steroid hormones. APs have been extensively studied in the past 25 years due to their importance from a medical and biological point of view, especially concerning their activity as oral contraceptives. In addition, the reproductive progesterone steroids and analogues are believed to be associated with a variety of cancers (breast, ovary, and endometrium).[1]

The attempt to rationalize the connection between the molecular structures of organic compounds and their biological activities constitutes the field of structure−activity relationship (SAR) studies. Correlation between structure and activity is important for the development of chemical agents (drugs, pesticides, etc.). These studies have practical importance because the results can be used to predict the activity of untested or hypothetical compounds. In addition, SAR studies can direct attention to molecular features that strongly correlate with biological activity, thus confirming or contradicting proposed mechanisms of action or suggesting further experiments.[2,3] Quantitative SARs (QSARs) are extensively used to correlate molecular structures of compounds with their biological activities, and many types of descriptors (experimental, measured, or computationally calculated) are used today.[4−7]

Among the most used SAR techniques we can certainly mention principal component analysis (PCA). Recently neural networks (NN) have also gained in popularity and have been employed as an additional tool in multivariate statistical analysis. Although widely used and in general with good results, their use in certain cases poses some inconveniences.[8] In PCA, for instance, it is very common to start using a large number of descriptors in order to select the best ones. This procedure could lead to descriptors with good statistical value but biochemically meaningless, thus limiting PCA analysis in terms of the elucidation of biochemical processes. On the other hand, NN predictions sometimes show a strong dependence on the choice of the training set, failing when more sophisticated tests such as cross-validation (leave-one-out) are used. Besides that, depending on the system studied, PCA and NN are very time-consuming. The search for simpler and sounder methodologies poses a continuous challenge in order to improve SAR studies.

Recently[9,10] a new methodology called EIM (electronic indices methodology) was proposed to investigate the structure−activity relationship of carcinogenic compounds. This methodology is based on the concept of local density of states[9] and critical values for the electronic energy difference among the frontier molecular orbital levels. By analyzing these quantum electronic descriptors, simple rules can be derived. The essence of such rules is logical conditional relations. Instead of purely statistical correlation (such as in PCA), EIM uses composed logical Boolean expressions. We like to stress that in fact the EIM approach is different from PCA or NN. EIM is an additional tool that can be applied in any kind of multivariate statistical analysis. In essence, EIM uses QSAR parameters to perform a qualitative SAR analysis through the use of discriminant analysis.

Initially EIM was successfully employed to describe structure−activity relationship of the carcinogenic activity of polycyclic aromatic hydrocarbons (PAHs).[9,10] EIM was applied to other classes of compounds, such as inhibitors of HIV integrase[11] and mytomicins[12] with very good results. Moreover, with relation to PAHs EIM was contrasted with

* Corresponding author. E-mail: rbraga@ifi.unicamp.br.

10.1021/ci000454f CCC: $19.00
Published on Web 09/26/2000

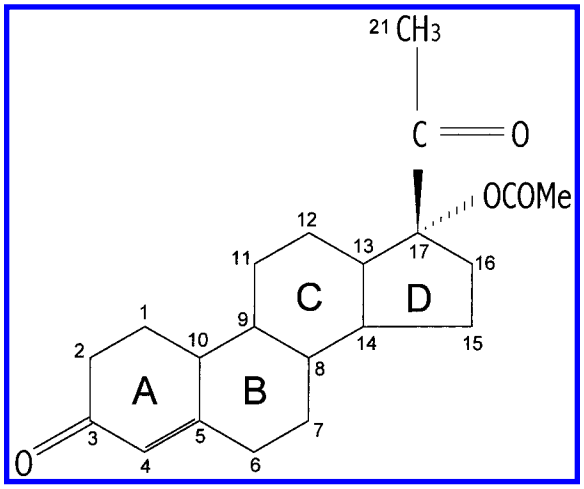

Figure 1. 17α-Acetoxyprogesterone showing steroidal fused ring skeleton and numbering (1−17) used internationally (IUPAC). Table 1 lists IUPAC names.

Table 1. Descriptive Names of the 21 Substituted 17α-Acetoxyprogesterones Studied[a]

no.	molecule	OPA	C
1	norethisterone	1	I
2	17α-acetoxyprogesterone	0.07	I
3	17α-ethinyltestosterone	0.20	I
4	21-chloro-1,6-bisdehydro-17α-acetoxyprogesterone	0.20	I
5	6α-nitro-17α-acetoxyprogesterone	0.21−0.28	I
6	6*β*-chloro-17α-acetoxyprogesterone	0.5	I
7	6α-fluoro-17α-acetoxyprogesterone	1	I
8	21-fluoro-1,6-bisdehydro-17α-acetoxyprogesterone	1	I
9	6α-bromo-17α-acetoxyprogesterone	1	I
10	6α-methyl-17α-acetoxyprogesterone	2−3	I
11	6α-chloro-17α-acetoxyprogesterone	2−3	I
12	6α-bromo-1-hydro-17α-acetoxyprogesterone	6	I
13	6α-fluoro-1-hydro-17α-acetoxyprogesterone	6	I
14	1,6-bisdehydro-6α-fluoro-17α-acetoxyprogesterone	8	I
15	1-hydro-6α-methyl-17α-acetoxyprogesterone	8	I
16	6α-chloro-1-hydro-17α-acetoxyprogesterone	8	I
17	6-methyl-1,6-bisdehydro-17α-acetoxyprogesterone	10	A
18	6-methyl-6-hydro-17α-acetoxyprogesterone	12	A
19	6-α-fluoro-6-hydro-17α-acetoxyprogesterone	15	A
20	6-chloro-1,6-bisdehydro-17α-acetoxyprogesterone	35	A
21	6-chloro-6-hydro-17α-acetoxyprogesterone	50	A

[a] See Figure 1 for their molecular structures. The oral progestational activities (OPA) relative to norethisterone[1] and the classification (C) are also indicated. A and I refer to active and inactive compounds, respectively.

PCA and NN studies[13] and produced the same order of predictive accuracy with fewer variables, simple rules, and significant reduction of computation efforts.

In this work we have applied EIM to the class of APs (compounds indicated in Figure 1 and Table 1). The main objective of this work is to investigate whether similar to other classes of organic compounds EIM methodology can be applied to identify active and inactive APs. Also, we would like to investigate the class of "universality" of the EIM parameters. Can they be used for any class of organic compounds?

We show that, similar to previous results with HIV integrase,[11] mitomycins,[12] and PAHs,[13] it is possible to correlate the progestational activity of APs with electronic indices through very simple rules. Moreover, we have also carried out PCA and NN calculations with the descriptors obtained from EIM. With less than the half of descriptors used before, we can predict biological activity with great accuracy (higher than 90%). These results suggest that the parameters from the EIM approach might be universal and certainly EIM is a very efficient/low-cost methodology for SAR studies.

2. METHODOLOGY

In the present work we have studied the same 21 17α-acetoxyprogesterones investigated before by one of us[8] using PCA/NN (Figure 1) in order to better contrast the results of the different techniques.

In relation to biological activity we have divided our molecules into only two groups: active and inactive (Table 1) according to their relative oral progestational activity.[1] This criterion has been proposed by Villemin and collaborators[14] and has been successfully used before. Following this criterion the molecules numbered from 1 to 16 are considered inactive (I) and the five remaining (17 to 21) are active (A).

As experimental geometric data for all the structures are not available, we have carried out fully geometric optimizations for the entire set of compounds. These calculations were performed using the well-known semiempirical method PM3 (parametric method 3[15]). The optimized geometries were obtained setting the gradient in the hypersurface of energy to be lower (in module) than 0.01 kcal/mol, to ensure good quality results. We have used the MOPAC program,[15] version 6.0. The total CPU time was less than 40 h in a Pentium PC, 400 MHz, with 256 MB of RAM memory.

Once the optimized geometries, eigenvalues, and eigenvectors were obtained, we proceeded to the calculation of density of states. The electronic density of states (DOS) is defined as the number of electronic states per energy unit. The related concept of local density of states (LDOS), i.e., the DOS calculated over a specific molecular region, is introduced in order to describe also the spatial distribution of the states over the system under consideration. For the LDOS calculations the contribution of each atom to an electronic level is weighted by the square of the (real) molecular orbital coefficient, i.e., by the probability density corresponding to the level in that site. The summation is carried over the selected atomic orbitals (n_I to n_f), leading to the expression[9−12]

$$\mathrm{LDOS}(E_i) = 2\sum_{m=n_i}^{n_f} |c_{mi}|^2 \qquad (1)$$

The factor 2 comes from the Pauli exclusion principle (maximum of two electrons per electronic level). This is a discrete modulation, which allows a direct comparison of DOS and LDOS calculated from any LCAO (linear combination of atomic orbital) method.

The EIM approach is based on two major descriptors, named η and Δ. η is related to the relative contribution difference between the most relevant molecular electronic levels over the identified molecular region linked to the biological activity (via LDOS):

$$\eta = 2\sum_{m=n_i}^{n_f} (|c_{m\mathrm{Level1}}|^2 - |c_{m\mathrm{Level2}}|^2) \qquad (2)$$

After an exploratory search we identified the region defined by the orbitals of the atoms 1−4 (see Figure 1) as the molecular region defining the biological activity within the EIM approach.

The LDOS analysis allows an easy identification of the relevant molecular levels correlated with biological activity. In the present work, we identified the highest occupied molecular orbital (HOMO) and its adjacent lower level (HOMO − 1) as the molecular levels mostly directly linked to biological activity. In this case

$$\eta = 2 \sum_{m=n_i}^{n_f} (|c_{m\mathrm{HOMO}}|^2 - |c_{m\mathrm{HOMO-1}}|^2) \qquad (3)$$

The second major EIM descriptor is Δ. This descriptor is defined as the energy separation of the molecular levels from eq 2, in this case the frontier levels HOMO and HOMO − 1. Using only these two descriptors, we are able to group and separate active and inactive APs.

The PCA and hierarchical cluster analysis (HCA) studies were carried out using the program package Pirouette,[16] which contains the PCA and related methods. The PCA presented results are for the best four descriptors obtained after a careful analysis over a set that included initially 17 descriptors. The calculated physicochemical descriptors used in the present work are the following: the highest occupied molecular orbital (HOMO) energy and its lower level (HOMO − 1); the energy difference between HOMO and HOMO − 1 (Δ energy); the HOMO, HOMO − 1 contribution (CH and CH − 1, respectively) to the local density of states (LDOS) over the A ring (Figure 1) and their difference η = (CH) − (CH − 1) (eqs 2 and 3).

The above quantum chemical descriptors were obtained directly from the PM3 calculations, and the reasons for their selection are discussed in details in refs 8−10.

Finally, the NN calculations were carried out using the program Perceptron-type Neural Network Simulator for Drug Design[17] (PSDD) (Quantum Chemistry Program Exchange No. 615, *http://qcpe.chem.indiana.edu/*). The back-propagation method was used and the calculation process was performed in a supervised way.

3. RESULTS AND DISCUSSION

In Table 2 we show a summary of the PM3 results for the 21 substituted APs. The values for the HOMO, HOMO − 1, and their energy difference (Δ) are indicated. There are also indicated the HOMO and HOMO − 1 LDOS contributions (CH and CH − 1) and their relative difference (η).

We have used the methodology developed for the PAHs as a guide for our present analysis of the AP data. Studying the same parameters (especially Δ and η) used with the PAHs, we observe from Table 2 distinct patterns for active and inactive APs. Using the EIM we can state the following rule 1 in order to classify the biological activity of the 17α-acetoxyprogesterones. This rule possesses exactly the same structure of the rule obtained before for other classes of compounds.[11−13] The η critical value is the same (0), but the Δ one changes.

Rule: If η < 0 and Δ > 0.800 eV, the molecule will be active; otherwise the molecule will be inactive. This simple rule correctly identifies the progestational activity of 90.5% of the molecules of Figure 2. Only the progestational activities of molecules **8** and **19** are incorrectly identified.

Table 2. Summary of PM3 Results for the 21 Substituted APs[a]

N	HOMO − 1	HOMO	Δ	CH − 1	CH	η
1	−10.4581	−10.1264	0.3317	0.6552	0.8817	0.2265
2	−10.5049	−10.2388	0.2661	0.6518	0.8703	0.2184
3	−10.4554	−10.1846	0.2708	0.6585	0.8712	0.2127
4	−10.5498	−9.5520	0.9978	0.3871	0.5829	0.1958
5	−10.8977	−10.7369	0.1608	0.7755	0.0144	−0.7611
6	−10.5732	−10.4533	0.1199	0.1316	0.6663	0.5347
7	−10.6295	−10.4910	0.1385	0.5491	0.8105	0.2615
8	−10.5596	−9.5556	1.0040	0.6417	0.5834	−0.0583
9	−10.6421	−10.4583	0.1838	0.2815	0.7874	0.5059
10	−10.4760	−10.2066	0.2694	0.6528	0.8693	0.2166
11	−10.6073	−10.4076	0.1997	0.6341	0.6511	0.0171
12	−10.6642	−10.4495	0.2147	0.5247	0.8398	0.3151
13	−10.6621	−10.4368	0.2254	0.4764	0.8687	0.3924
14	−10.6857	−9.7042	0.9814	0.0435	0.4589	0.4155
15	−10.5019	−10.2077	0.2943	0.6121	0.8639	0.2518
16	−10.6433	−10.3822	0.2611	0.5747	0.7283	0.1536
17	−10.5411	−9.4151	1.1261	0.6308	0.4695	−0.1613
18	−10.5061	−9.3918	1.1143	0.6633	0.4927	−0.1705
19	−10.6636	−9.6652	0.9984	0.0117	0.4713	0.4596
20	−10.5778	−9.4724	1.1054	0.7440	0.3278	−0.4161
21	−10.5810	−9.4453	1.1356	0.5959	0.3380	−0.2579

[a] The values for the HOMO (H), HOMO − 1 (H − 1), and their energy difference (Δ) are indicated. There are also indicated the HOMO (CH) and HOMO − 1 (CH − 1) LDOS contribution and their relative difference (η). The LDOS was carried out over the region 1−4 indicated in Figure 1. Energies are expressed in EV, and the relative contributions are expressed in the normalized population charge values (from 0 to 2). The descriptors 3 and 6 (Δ and η) were also used in the electronic index methodology.[9−13]

The AP data from Table 2 present some features that were not observed with the class of the PAHs. When we analyze separately the parameters η and Δ, we observe that they individually contain enough information that can be used to classify the APs with a high accuracy (this was not possible with the PAHs). As can be seen from Table 2, the molecules numbered **1**−**16** (inactive) have in general Δ values between 0.12 and 0.332 eV, and the molecules numbered **17**−**21** (active) Δ values between 0.998 and 1.136 eV. Molecules **4**, **8**, and **14** constitute exceptions: they are inactive but present Δ of the order of 1 eV. The existence of higher Δ related to higher progestational activity is a result new to this class of compounds. This never had been speculated before. It seems that "clean" frontier orbital, i.e., a HOMO well separated in energy from HOMO − 1 is a necessary (but not sufficient) condition for active hormones. This result was also observed with active carcinogenic PAHs.[9,10]

The determination of ring A as the relevant region was obtained from exploratory analysis of the LDOS over many limited geometrical regions of the substituted APs. We carried out LDOS calculations over the relevant positions (3, 6, 10, and 17 of the steroid skeleton, see Figure 1) believed to be important.[8] Our results for the LDOS from these regions did not provide patterns that could be correlated with the progestational activity. The same was observed for the LDOS involving the terminal ring D, the intermediate rings C and B (Figure 1), and other molecular regions. However, when we analyzed the LDOS on positions 1, 2, 3, and 4 simultaneously (ring A), a clear pattern appeared (see η values in Table 2) that lead us to the rule stated above. In general molecules **1**−**16** (inactive) present η > 0, except for molecules **5** and **8**. Molecules **17**−**21** (active) exhibited

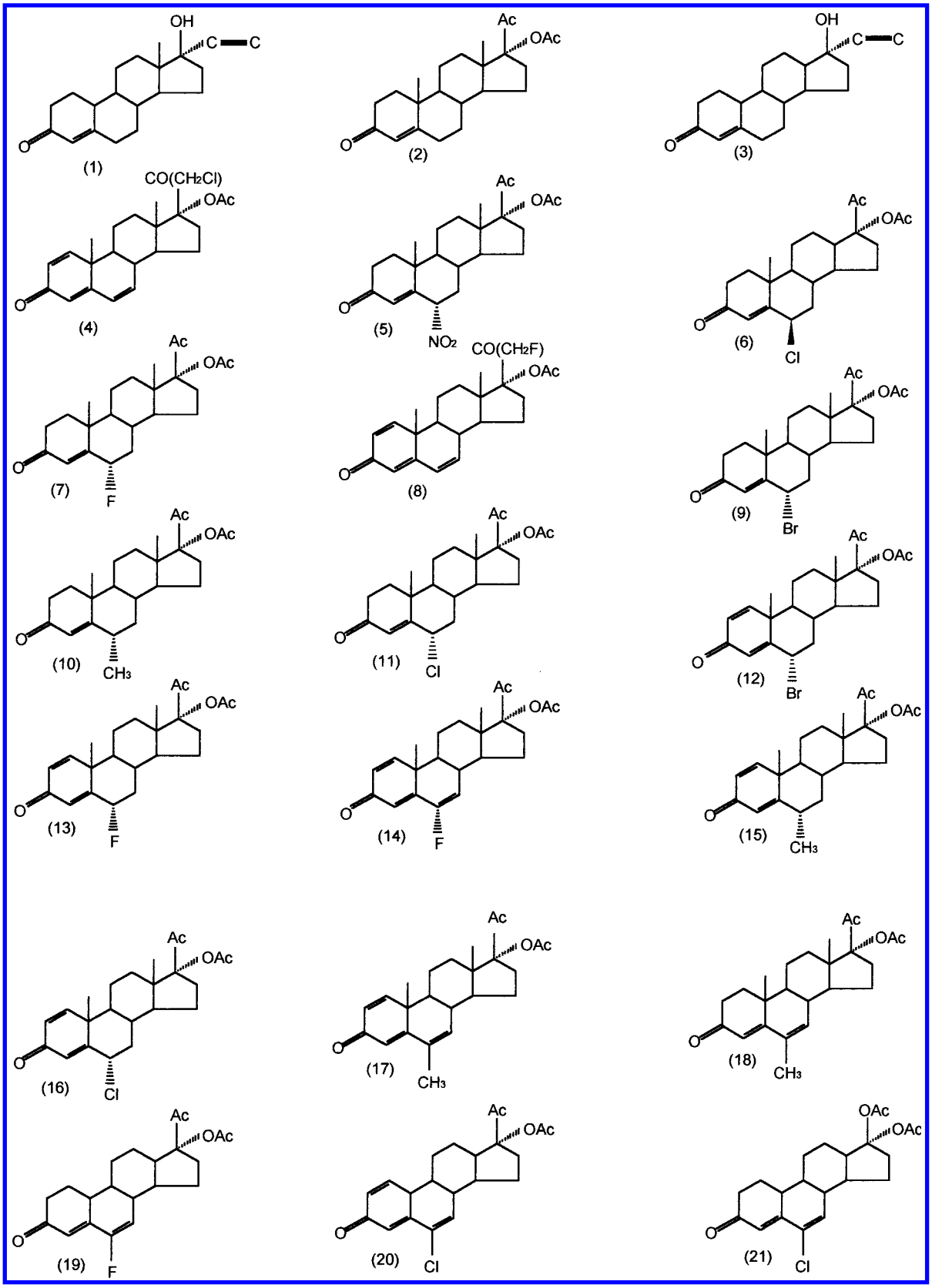

Figure 2. Molecular structure of the 21 17α-acetoxyprogesterone molecules. Table 1 lists their IUPAC names.

$\eta < 0$, except for molecule **19**. This indicates that the combination of the relative density of states (η) and the difference from the frontier states (Δ) provides an important quantum chemical descriptor of biological activity.

The use of DOS and LDOS concepts can give us detailed information on the contributions of specific geometrical regions of the molecules to the chemical reactivity, optical response, etc. In the case of substituted APs, only ring A presented patterns that can be associated with the progestational activity. In fact, ring A plays an important role in the steroid and the receptor interactions. A progestogen receptor site establishes close specific contact only with ring A.[18] The

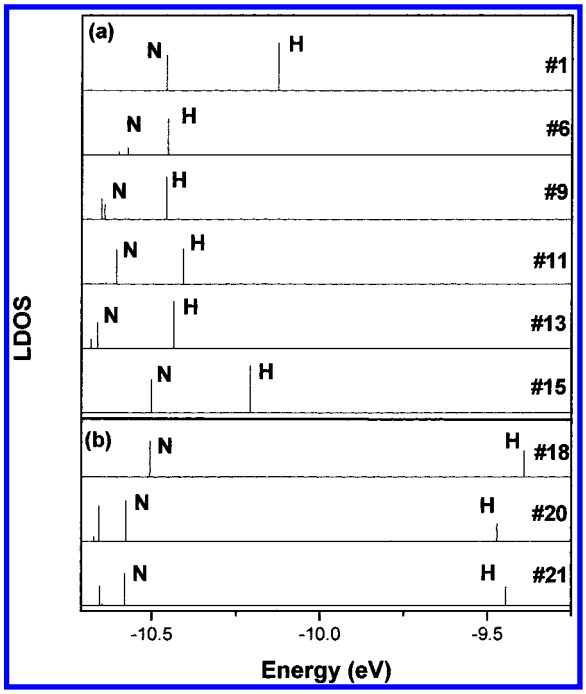

Figure 3. Local density of states (LDOS) over the positions 1, 2, 3, and 4 (see Figure 1) for substituted 17α-acetoxyprogesterones representative of the rules stated in the text. H indicates the highest occupied molecular orbital (HOMO) and N its next one (HOMO − 1).

progestogen receptor sites that interact with ring A of the steroid are well-known.[19] The patterns presented by ring A related to the biological activity of substituted APs reflect the chemical reactivity presented by this molecular region.

In Figure 3 we show the LDOS on positions 1, 2, 3, and 4 (ring A) for two typical kinds of results. In part a we show the results for the inactive molecules **1**, **6**, **9**, **11**, **13**, and **15**. We have $\Delta < 0.800$ eV and the HOMO contribution greater than that of HOMO − 1, and according to our rule, they should be inactive. This is in agreement with the experimental data. Similarly, in part b we show the results for the active molecules **18**, **20**, and **21**. These molecules have $\Delta > 0.800$ eV and the relative HOMO − 1 contribution to the LDOS greater than that of HOMO, and consequently, according to our rule, they should be active. This is again in agreement with the experimental data. Moreover, we observed that the active molecules present not only high values of Δ but also a significant blue shift of the frontier levels, when we compare them to the inactive ones.

These results are very surprising due to the fact that they are very similar to the ones obtained in our previous works with the class of PAHs.[9,10,13] It looks like the EIM parameters possess some universality in classifying the biological activity (active vs inactive). Although these classes have very different biological behaviors (hormones and carcinogens), it is intriguing that the EIM approach points to the same descriptors separating active and inactive compounds.[13] This suggests that some common topological features might be present and deserves more detailed investigation.

If the descriptors pointed out by the EIM are relevant to APs, we can expect that sophisticated SAR techniques will lead naturally to them. Also, with a basis in previous EIM studies we can expect that the use of the indices that arise from EIM will improve previous SAR results.[8] To test these hypotheses, we performed PCA and NN analysis using some of the quantum chemical descriptors of Table 2.

The first molecular set studied with the PCA method were molecules **1**−**9** and **17**−**21** shown in Figure 2. This is the same set of compounds used before in the literature,[8] in order to allow a direct comparison between the different methodologies. From the descriptor list mentioned in the methodology section, the best separation in active and inactive compounds was obtained using the following ones (Table 2): HOMO, Δ, CH, and CH − 1.

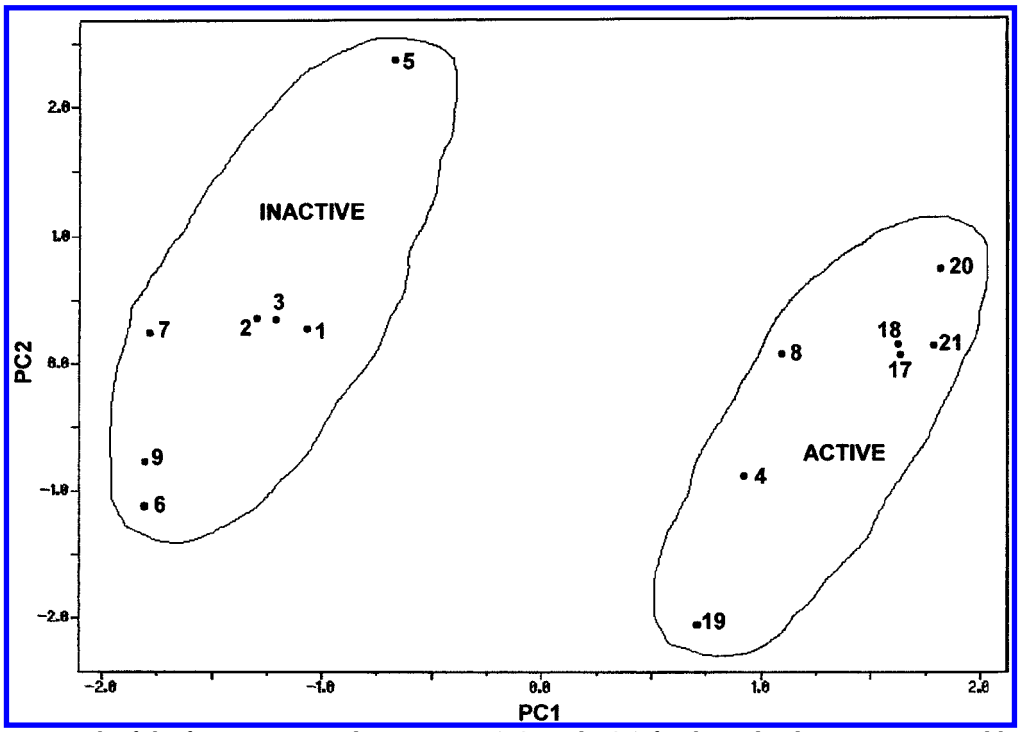

Figure 4. The score graph of the first two principal components (PC1 and PC2) for the molecular training set used before[2] (molecules 1−9, 17−21).

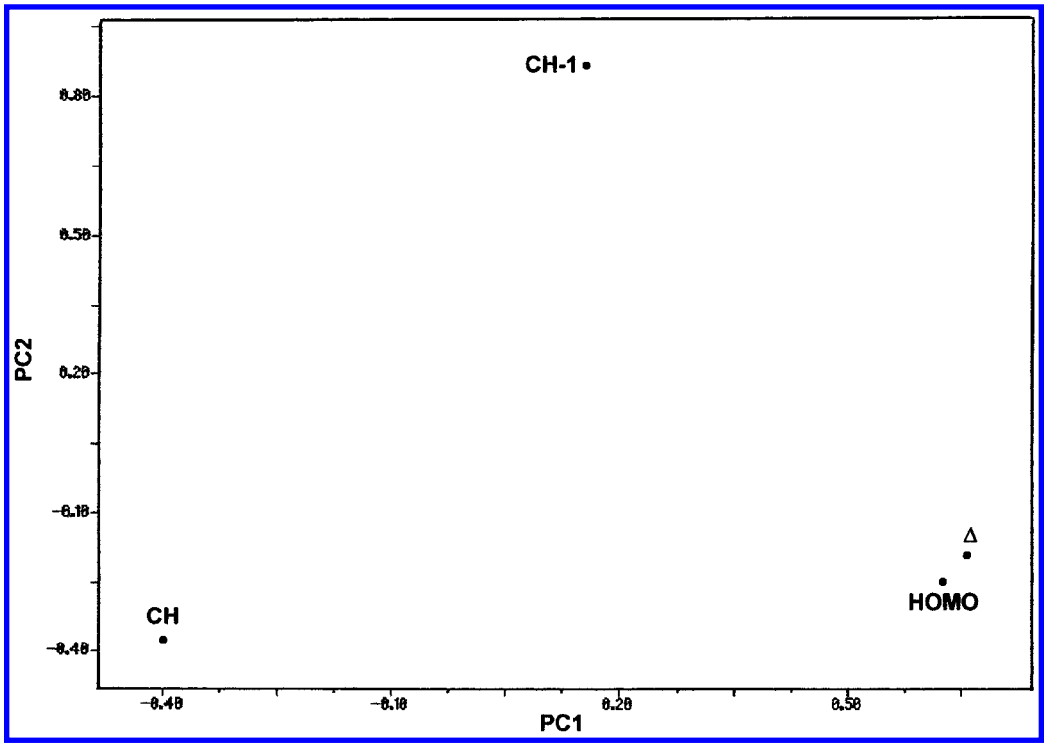

Figure 5. Loadings of the four physicochemical descriptors (HOMO, CH, CH − 1, and Δ, described in Table 3) selected for PCA to the set of molecules studied.

In Figure 4 we show the scores of the first two principal components (PC1 and PC2) for this first set of 14 hormones. The molecules are distributed into two distinct regions in this figure. The active group is on the right side and the inactive one on the left side. Molecules **4** and **8** are incorrectly classified as active. Twelve molecules out of the 14 are correctly classified, which corresponds to an accuracy of 85.7%.

The two principal components (PC1 and PC2) are given by

$$\mathrm{PC1} = 0.62(\mathrm{HOMO}) + 0.66(\Delta) + 0.15(\mathrm{CH} - 1) - 0.40(\mathrm{CH}) \quad (4)$$

$$\mathrm{PC2} = -0.25(\mathrm{HOMO}) - 0.19(\Delta) + 0.87(\mathrm{CH} - 1) - 0.38(\mathrm{CH}) \quad (5)$$

PC1 and PC2 respond to 55.35% and 25.93% of the variance, respectively. Equation 1 indicates that HOMO and the difference in energy between HOMO and HOMO − 1 (Δ) are the major descriptors of PC1. The major descriptor of PC2 is the HOMO − 1 contribution to the LDOS (CH − 1). It is interesting to note that the highest contributions for PC1 and PC2 are exactly the same most important variables derived from EIM methodology.

In Figure 5 we show loadings of the four physicochemical descriptors. The descriptors are grouped in three regions: at the left side of this figure, with the values of the PC1 axis less than −0.2 (descriptor CH), in the upper center with the values of the PC1 axis greater than 0.8 (descriptor CH − 1), and on the bottom of the right side (descriptors HOMO and Δ). A comparison between Figures 4 and 5 shows that the two descriptors HOMO and Δ are responsible for pulling the active molecules toward to the right side in the score graph (Figure 4). The descriptors that are mostly responsible for pulling inactive molecules toward the left side and toward the upper center in Figure 4 are CH and CH − 1, respectively.

To study the predictive ability of the PCA method, we applied PCA to the remaining 7 molecules (**10**−**16** in Figure 2), using exactly the same four descriptors as those used for the set of 14 hormones (Figures 4 and 5). The score graph of PCA is illustrated in Figure 6. The active group is located on the right side of this figure and the inactive one on the left side. Molecules **4**, **8**, and **14** are incorrectly classified as actives. Out of the seven newly added compounds, six molecules are correctly classified, which corresponds to an accuracy of 85.7% in the prediction.

The two principal components (PC1 and PC2) for the entire set of hormones are given by

$$\mathrm{PC1} = 0.612(\mathrm{HOMO}) + 0.645(\Delta) - 0.067(\mathrm{CH} - 1) - 0.453(\mathrm{CH}) \quad (6)$$

$$\mathrm{PC2} = -0.001(\mathrm{HOMO}) - 0.026(\Delta) + 0.983(\mathrm{CH} - 1) - 0.183(\mathrm{CH}) \quad (7)$$

PC1 and PC2 respond to 58.16% and 25.15% of the variance, respectively. Again we note that the highest contributions for PC1 and PC2 are exactly the same most important variables derived from EIM methodology[9,10] (eqs 4 and 5). Eighteen molecules out of the global set of 21 APs are correctly classified, which also corresponds to 85.7% overall correct classification, slightly inferior to the EIM's results (90.5%).

Figure 7 shows the hierarchical clustering dendogram. Hierarchical clustering dendogram (Hier) is a conceptually simple but effective clustering technique. Given a set of compounds and a list of descriptors, this technique clusters the compounds that share a common property (similarity). See refs 14, 20, 21, and 22 for details about dendogram techniques. In the present case the methodology used for

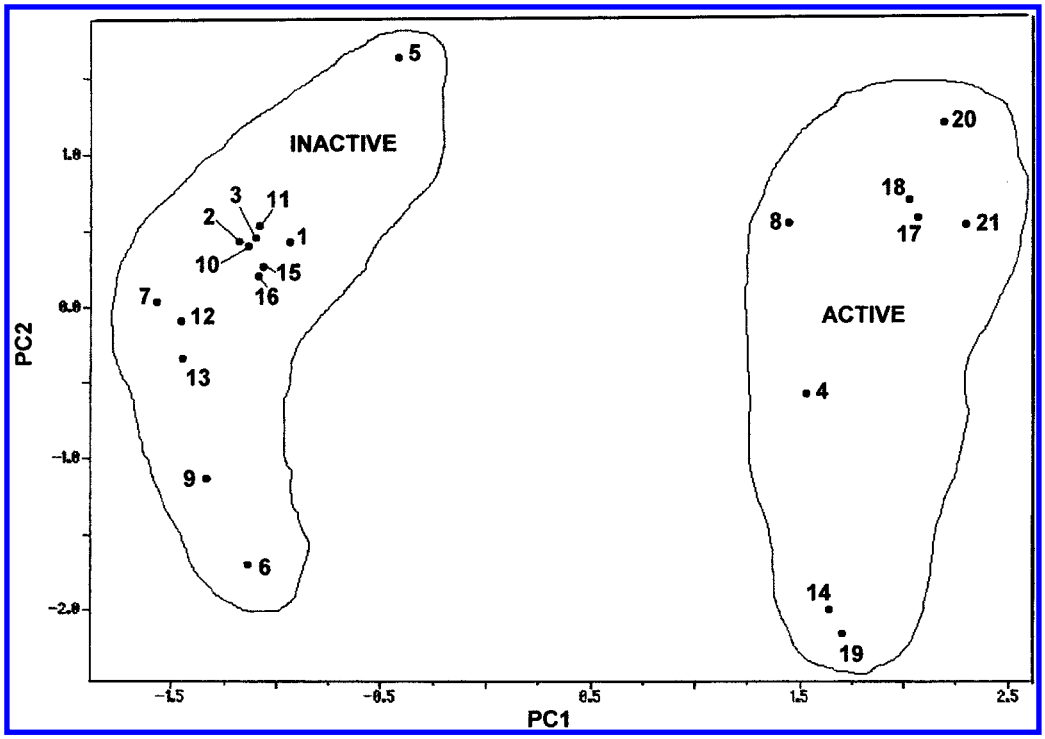

Figure 6. The score graph of the first two principal components (PC1 and PC2) for the global set of molecules (Figure 2).

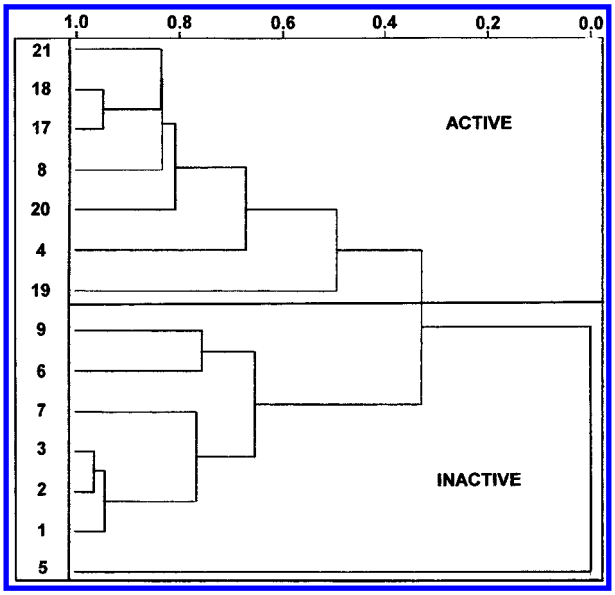

Figure 7. Dendogram obtained for the training set (molecules 1−9, 17−21) with the physicochemical descriptors HOMO, CH, CH − 1, and Δ, described in Table 3. The darker line is to aid in the visualization of the active and inactive clusters.

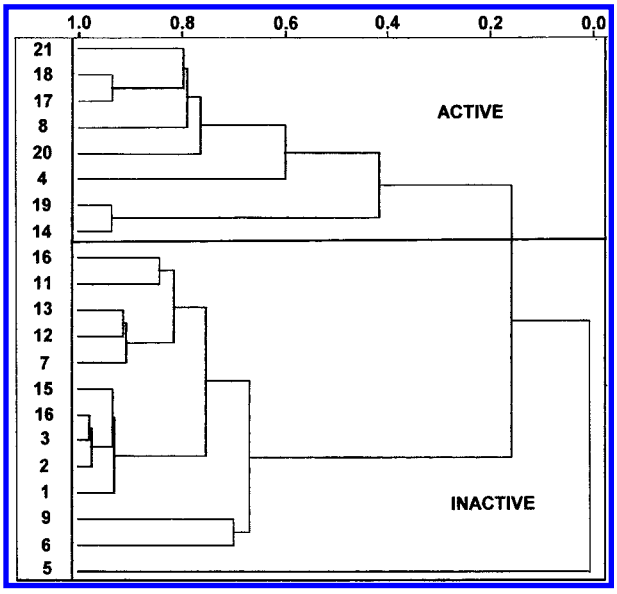

Figure 8. Dendogram obtained for the entire set of molecules (see Figure 2) with the physicochemical descriptors HOMO, CH, CH − 1, and Δ, described in Table 3. The darker line is to aid in the visualization of the active and inactive clusters.

the hierarchical clustering is a simple Euclidean distance, followed by a standard similarity index computation.

From the dendogram (Figure 7) we can see that with the use of four descriptors in the PCA calculations two clusters are formed (separated by the horizontal line between molecules **19** and **9**) with the same training set used in PCA calculations. One group is mostly composed of active molecules (upper half of the figure) and the other by inactive molecules (lower half). These two clusters have zero similarity. This demonstrates that active and inactive molecules are well separated in this four-dimensional space. The molecules incorrectly classified using Hier are the same as those observed with the PCA analysis.

We show in Figure 8 the dendogram for the entire set of hormones. We have again two distinct clusters of active/inactive molecules, with zero similarity. As in PCA, the molecules incorrectly classified were **4**, **8**, and **14**.

In summary, with the descriptors pointed out by EIM, we could introduce an overall simplification of the classifications of the class of APs with the same accuracy as before.[8] The global results for the final rule of EIM and for the PCA analysis were 90.5% and 85.7%, respectively.

Next, NN was employed to study SAR of the APs. For these studies we adopted the following procedure. We started studying the 12 AP molecules (**2**−**9**, **17**−**19**, and **21**) using four neurons in the first layer that are the same four

Table 3. Parameters Used (α, θ, and ϵ) in the Neural Network Calculations[a]

layer	neurons	α	θ	ϵ
1	4			
2	13	16.0	0.0	0.001
3	2	12.5	0.0	0.1

[a] α is the nonlinear parameter of the sigmoid functions, θ is a threshold value for a neuron,[23] and ϵ is a parameter which determines the shift for correction in recursive cycles. The NN have used 1.078 training epochs.

Table 4. Results from the Neural Network (NN) Calculations[a]

molecule	category	training pattern	output pattern
2	1	1 0	1.000 0.000
3	1	1 0	1.000 0.000
4	1	1 0	0.979 0.074
5	1	1 0	1.000 0.000
6	1	1 0	1.000 0.000
7	1	1 0	1.000 0.000
8	1	1 0	0.998 0.038
9	1	1 0	1.000 0.000
17	2	0 1	0.007 0.999
18	2	0 1	0.034 0.995
19	2	0 1	0.038 0.956
21	2	0 1	0.000 1.000

[a] NN trained with 12 progesterone molecules (**2**−**9**, **17**, **18**, **19**, and **21** of Figure 1) using the following parameters: HOMO, Δ, CH − 1 and CH.

Table 5. Prediction of the Progestational Activity for Nine Progesterones: Molecules **1**, **10**−**16**, and **20** (Figure 1), Using the NN Trained in Table 3; Error = 0.000 000 0

molecule	category	predicted pattern
1	1	1.000 0.000
10	1	1.000 0.000
11	1	1.000 0.000
12	1	1.000 0.000
13	1	1.000 0.000
14	2	0.054 0.945
15	1	1.000 0.000
16	1	1.000 0.000
20	2	0.000 1.000

physicochemical descriptors used in PCA, in order to allow a direct comparison with the PCA results. The parameters used in NN calculations are listed in Table 3. The perceptron-type NN consists of three layers. The training of NN was carried out according to the back-propagation algorithm[22] until the error function[23] reached the convergence criterion ($\leq$0.000 01 in our case).

α is a parameter which expresses the nonlinearity of the neuron's operation.[23] It is a value of the sigmoid function of the neurons in the second and third layers. Its default value is 1.0. α was forced to change (according to the values in Table 3) until the error function reached the convergence. θ is a threshold value for the neuron in the second and third layers.[23] It was set to its usual value (0.0).

We trained the NN with the 12 APs mentioned above to predict progestational activity of the remaining compounds. Table 4 shows the results of NN training and Table 5 show the prediction results. The inactive group belongs to category 1 and the active one belongs to category 2. The training pattern of category 1 is (1 0), whereas the training pattern of category 2 is (0 1) as seen in Table 5. In the initial phase of "NN learning", the weight matrix was calculated with the training pattern using the four parameters for each of the 12 molecules. Although the neurons at the first layer can take continuous values between 0 and 1, those of the last layer (output patterns) are required to assume discrete values of 0 or 1. However, in general, the final output does not present a complete set of discrete values. This is due to the limit of resolution ability of the network, but it is expected that these values would be close to the discrete ones [(1 0) or (0 1)]. The classification criterion in these cases is based on its closer proximity to these limits [for instance (0.998 0.002) to (1 0), and (0.004 0.996) to (0 1)]. The NN "learned" the training pattern with success (100%). Molecules **4** and **8**, incorrectly classified with PCA, are now correctly described.

For the remaining nine molecules, as we can see in Table 5, the NN correctly predicts the progestational activity for eight of them, which corresponds to an accuracy of 88.9%. A total of 20 molecules out of the global set of 21 APs are correctly classified, which corresponds to 95.2% correct global classification.

Tables 6 and 7 compare the performance of our three different methods (EIM, PCA, and NN) for the studied compounds. These results refer to PCA and NN data sets using the same four descriptors mentioned above compared to EIM ones using only two of them. On average all three methods have approximately the same margin of correct

Table 6. Summary of the Theoretical Predictions Contrasted to the Experimental Data[a]

	rule 1	rule 2	rule 3		
N	Δ	η	Δ and η	PCA	NN
1	+	+	+	+	+
2	+	+	+	+	+
3	+	+	+	+	+
4	−	+	+	−	+
5	+	−	+	+	+
6	+	+	+	+	+
7	+	+	+	+	+
8	−	−	−	−	+
9	+	+	+	+	+
10	+	+	+	+	+
11	+	+	+	+	+
12	+	+	+	+	+
13	+	+	+	+	+
14	−	+	+	−	−
15	+	+	+	+	+
16	+	+	+	+	+
17	+	+	+	+	+
18	+	+	+	+	+
19	+	−	−	+	+
20	+	+	+	+	+
21	+	+	+	+	+
%	85.7	85.7	90.5	85.7	95.2

[a] + and − mean in agreement/disagreement with the experimental data. Results for rules 1, 2, and 3 stated in the text, and from the PCA and NN analysis. There are also indicated the percentage of corrected predictions for each procedure. The PCA and NN calculations were carried out with the following descriptors: HOMO, Δ, CH − 1, and CH.

Table 7. Percentage of Correct Classification for the Three Different Methodologies (EIM, PCA, and NN) for the Entire Set of Molecules (Figure 2)

methodology	performance
EIM	90.5%
PCA	85.7%
NN	95.2%

prediction (90%). There are, however, slight differences in the percentages attained among the three different methods. NN seems to give slightly better percentage than EIM, which gives slightly better percentage than PCA.

4. SUMMARY AND CONCLUSIONS

In this work we applied the electronic indices methodology to study the biological activity of the class of 17α-acetoxyprogesterones. We have studied 21 compounds (Figures 1 and 2), and we could correctly describe the active/inactive compounds with an accuracy of 90.5%, using the parameters from EIM. Using these same parameters, we have improved previous SAR studies,[8] with fewer descriptors. Our principal component analysis and neural network approaches correctly describe this set of hormones with an accuracy of 85.7% and 95.2%, respectively. Moreover, the PCA analysis naturally pointed out that the EIM parameters are the most important in the classification of this class. The combination of the EIM, PCA, and NN techniques provides an improvement of the classification results through a low computational cost.

The above results show the importance of the electronic indices Δ and the contribution to the LDOS on the frontier orbitals (CH and CH − 1) in the description of the biological activity of the hormones studied here. Moreover, a clear pattern appears when we proceed with the calculation of the LDOS over the A ring (Figure 1), reflecting its chemical importance.

Besides the present AP study, we have applied EIM methodology to methylated and nonmethylated PAHs,[8−10] HIV protease inhibitors,[11] mitomycins,[12] and taxoids.[24] These are different classes of organic compounds with very differentiated biological activities, and the same EIM descriptors were able to successfully classify active and inactive compounds.

These results reinforce the general applicability of the EIM approach, but it remains to elucidate why this occurs. In principle these quantum descriptors can construct geometric spaces which have a natural dipheomorphism with chaotic domains. Therefore the fact that we are using the same descriptors seems to be a topological feature correlated with the interactions with the DNA. The tentative correlation of geometric/electronic features with topological DNA properties has been intensively explored in past years,[25] and we believe that the EIM studies will add some useful information to these investigations. Studies through spaces which possess natural dipheomorphism with the spaces generated by these parameters are in progress,[26] and preliminary results indicate the presence of common fractal signatures for all of them.

Finally, we observe that the EIM parameter in connection with SAR studies could provide a new tool to optimize the study of many different classes of organic compounds. A software version of the EIM programs[27] can be freely obtained from our Web page: http://www.ifi.unicamp.br/gsonm/chem2pac/.

ACKNOWLEDGMENT

We thank Profs. P. M. V. B. Barone and Y. Takahata for helpful discussions. The authors also thank the Brazilian agencies FAPESP and CNPq for the financial support.

REFERENCES AND NOTES

(1) Shoppee, C. W. *Chemistry of the Steroids*, 2nd ed.; Spottiwoode, Ed.; Ballantyne & Co. Ltd.: London, 1964, and references therein.
(2) Jurs, P. C.; Chou, J. T.; Yuan, M. *J. Med. Chem.* **1979**, *22*, 476.
(3) Jurs, P. C.; Stouch, T. R.; Czerwinski, M.; Narvaez, J. N. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 296.
(4) Sutter, J. M.; Peterson, T. A.; Jurs, P. C. *Anal. Chim. Acta* **1997**, *342*, 113.
(5) Karelson, M.; Lobanov, V. S.; Katritzky, A. R. *Chem. Rev.* **1996**, *96*, 1027.
(6) Wilson, L. Y.; Famini, G. R. *J. Med. Chem.* **1991**, *34*, 1668.
(7) Politzer, P.; Murray, J. S.; Flodmark, P. *J. Phys. Chem.* **1996**, *100*, 5538.
(8) Vendrame, R.; Takahata, Y. *J. Mol. Struct. (THEOCHEM)* **1999**, *489*, 55.
(9) Barone, P. M. V. B.; Camilo, Jr., A.; Galvão, D. S. *Phys. Rev. Lett.* **1996**, *77*, 1186.
(10) Braga, R. S.; Barone, P. M. V. B.; Galvão, D. S. *J. Mol. Struct. (THEOCHEM)* **1999**, *464*, 257.
(11) Cyrillo, M.; Galvão, D. S. *J. Mol. Struct. (THEOCHEM)*, **1999**, *464*, 267.
(12) Santo, L. L. E.; Galvão, D. S. *J. Mol. Struct. (THEOCHEM)*, **1999**, *464*, 273.
(13) Vendrame, R.; Braga, R. S.; Takahata, Y.; Galvão, D. S. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1094.
(14) Villemin, D.; Cherqaoui, D.; Mesbah, A. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1288.
(15) Stewart, J. J. P. *J. Comput. Chem.* **1991**, *10*, 209; *J. Comput. Chem.* **1991**, *10*, 221; *Mopac Program*, version 6.0; Quantum Chemistry Exchange No. 455.
(16) *Pirouette Multivariate Data Analysis for IBM PC Systems*, Version 2.0; Infometrix: Seattle, WA, 1996.
(17) Ichikawa, H. PSDD: Perceptron-type Neural Network Simulator; QCPE 615; http://qcpe.chem.indiana.edu/.
(18) Duax, W. L.; Griffin, J. F.; Weeks, C. M. In *Interaction of Steroid Hormone Receptors with DNA*; Sluyser, M., Ed.; Ellis Horwood Ltd.: Chichester, England, 1985; p 83.
(19) Carlstedt-Duke, J.; Strömstedt, P. E.; Persson, B.; Cederlund, E.; Gustafsson, J. A.; Jörnvall, H. *J. Biol. Chem.* **1988**, *263*, 6842.
(20) Jyrol, R. C.; Bailey, D. E. *Cluster Analysis*; McGraw-Hill: New York, 1970.
(21) Kowalski, B. R.; Bender, C. F. *J. Am. Chem. Soc.* **1972**, *94*, 5632.
(22) Rumelhart, D. E.; McClelland, J. L. *Parallel Distributed Processing Exploration in Microstructure of Cognition.* MIT Press: Cambridge, MA, 1988.
(23) Aoyama, T.; Suzuki, Y.; Ichikawa, H. *J. Med. Chem.* **1990**, *33*, 2583.
(24) Braga, S. F.; Galvão, D. S. Unpublished.
(25) Bashford, J. D.; Tsohantjis, I.; Jarvis, P. D. *Proc. Natl. Acad. Sci. U.S.A.* **1998**, *95*, 987.
(26) Braga, R. S.; Galvão, D. S. Unpublished.
(27) Cyrillo, M.; Galvão, D. S. *EPA Newslett.* **1999**, *67*, 31.

CI000454F