A Structure−Activity Study of Taxol, Taxotere, and Derivatives Using the Electronic Indices Methodology (EIM)

S. F. Braga* and D. S. Galvão

Instituto de Física Gleb Wataghin, Universidade Estadual de Campinas,
C.P. 6165, CEP 13083-970, Campinas -S.P., Brasil

Received November 21, 2002

Among the new families of effective anticancer drugs, the natural product paclitaxel (Taxol/Bristol-Myers-Squibb) and its semisynthetic derivative docetaxel (Taxotere/Rhone-Poulenc Rorer) are probably the most promising agents under investigation. Surprisingly considering their importance no detailed quantum mechanical studies have been carried out for these drugs. In this work we report the first structure−activity relationship (SAR) studies for 20 taxoid structures using molecular descriptors from all-electron quantum methods. The used methods were the pattern-recognition Principal Component Analysis (PCA), Hierarchical Clustering Analysis (HCA), and the recently developed Electronic Indices Methodology (EIM). The combined use of EIM with PCA/HCA methodologies was able to correctly classify active and inactive taxoids with 100% of accuracy using only a few “universal” quantum molecular descriptors. It was possible to identify the electronic features defining active molecules. This information can be used to select and design new active compounds. The combined use of EIM with PCA/HCA can be a new and very efficient tool in the field of computer assisted drug design.

INTRODUCTION

Cancer is a general term regarding hundreds of related diseases. It is believed that it is consequence of a combination of complex factors related to heredity, lifestyle, and environment.[1−3] Cancers are characterized by an accelerated cellular growth accumulating genetic alterations as they progress to a more malignant phenotype,[4] and are nowadays the second major cause of human death in industrialized countries.[5]

These death rates were significantly reduced in the past decades due in part to the successful use of chemotherapeutic agents.[3] Among the new families of drugs, the natural product paclitaxel (Taxol/Bristol-Myers-Squibb) (Figure 1) and its semisynthetic derivative docetaxel (Taxotere/Rhone-Poulenc Rorer) (Figure 2) are probably the most promising antitumoral agents under investigation.[6]

Paclitaxel was first isolated in minute quantities from the extract of the inner bark of the Pacific Yew tree *Taxus Brevifolia*, and its structure and activity against a number of tumors was identified in 1971 by Wall and his collaborators.[7] Docetaxel was obtained as a synthetic product in 1981 during a semisynthetic route[8] proposed to paclitaxel from 10-deacetilbaccatin III, a natural substance isolated from the leaves of the yew *Taxus Baccata*.[9] Although paclitaxel total synthesis is available, it is extremely complex to manufacturing production.[10]

These two antitumoral compounds have a unique mechanism of action. They block cancer cell division by promoting the polymerization and stabilization of tubulin protein heterodimers to microtubules and they also inhibit the depolymerization of microtubules back to tubulin.[11]

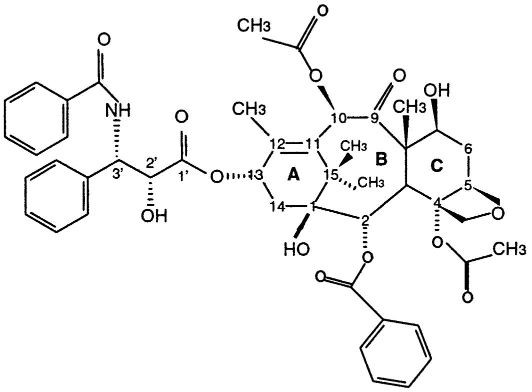

Figure 1. Structural representation of paclitaxel (Taxol).

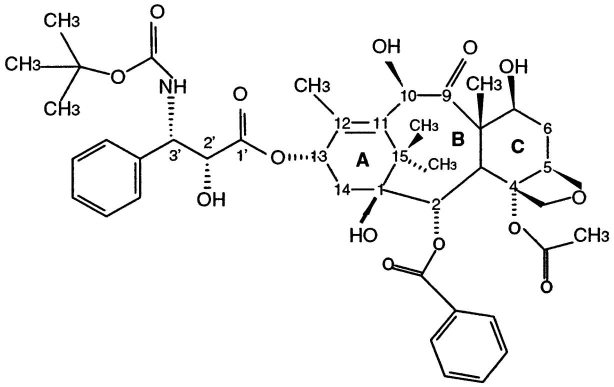

Figure 2. Structural representation of docetaxel (Taxotere).

The very complex structure of paclitaxel and docetaxel derivatives, termed taxoids, is responsible for the reduced number of X-ray studies obtained in three decades of research.[12,13] Subsequent studies about the three-dimensional

* Corresponding author phone: +55-19-3788-5345; fax: +55-19-3788-5376; e-mail: scheila@ifi.unicamp.br.

10.1021/ci025640v CCC: $25.00
Published on Web 02/27/2003

Table 1. 20 Investigated Taxoids[b]

molecules	R_2	R_3	R_4	ID_{50}/ID_{50}(paclitaxel)[a]	Q.A.
TX1-A (2′R, 3′S)	$COCH_3$	OH	$NHCO_2tBu$	0.5	A
TX1-B (2′S, 3′R)	$COCH_3$	OH	$NHCO_2tBu$	30	I
TX2-A (2′R, 3′S)	$COCH_3$	$NHCO_2tBu$	OH	10	I
TX2-B (2′S, 3′R)	$COCH_3$	$NHCO_2tBu$	OH	108	I
TX3-A (2′R, 3′S)	$COCH_3$	OH	NHCOPh	1	A
TX3-B (2′S, 3′R)	$COCH_3$	OH	NHCOPh	4.5	A
TX4-A (2′R, 3′S)	$COCH_3$	NHCOPh	OH	10	I
TX4-B (2′S, 3′R)	$COCH_3$	NHCOPh	OH	110	I
TT1-A (2′R, 3′S)	H	OH	$NHCO_2tBu$	0.5	A
TT1-B (2′S, 3′R)	H	OH	$NHCO_2tBu$	30	I
TT1-C (2′R, 3′R)	H	OH	$NHCO_2tBu$	1.8	A
TT1-D (2′S, 3′S)	H	OH	$NHCO_2tBu$	4.3	A
TT2-A (2′R, 3′S)	H	$NHCO_2tBu$	OH	10	I
TT2-B (2′S, 3′R)	H	$NHCO_2tBu$	OH	160	I
TT3-A (2′R, 3′S)	H	OH	NHCOPh	1.3	A
TT3-B (2′S, 3′R)	H	OH	NHCOPh	4	A
TT3-C (2′R, 3′R)	H	OH	NHCOPh	1.3	A
TT3-D (2′S, 3′S)	H	OH	NHCOPh	1.3	A
TT4-A (2′R, 3′S)	H	NHCOPh	OH	10	I
TT4-B (2′S, 3′R)	H	NHCOPh	OH	170	I

[a] ID_{50}: drug concentration that decrease in 50% the microtubules depolymerization (mM). [b] *A, B, C,* and *D* index represent compounds with 2′R, 3′S, 2′S, 3′R, 2′R, 3′R, and 2′S, 3′S stereochemistry, respectively. The qualitative activity (Q.A.) of the taxoids are also presented according to our classification of active (A) and inactive (I) molecules. See text for discussions.

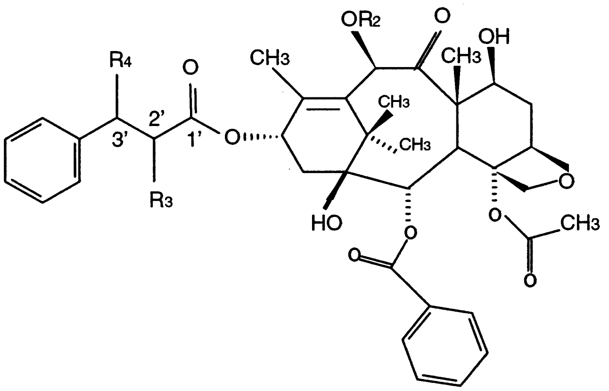

Figure 3. Structural representation of the studied taxoids. The radicals (Rn) and the chain carbon stereochemistry are indicated in Table 1.

structure of solvated paclitaxel-like compounds have showed strong solvent dependence on the conformation of phenylisoserine side chain crystallization.[14−18]

A large number of experimental investigations highlighted that the presence or absence of side-groups at some specific regions of the taxane diterpenoid skeleton and their stereochemistry are essential to the activity of paclitaxel analogues.[6,19−24]

Surprisingly, considering the taxoids importance, no detailed theoretical quantum structure−activity relationship (SAR) studies have been reported in the literature. This could be attributed in part to the size and conformational complexity presented by taxoids.

In this work we present the first systematic investigation of electronic and quantum features for a series of 20 taxoids (Figure 3, Table 1) for which the in vitro antimitotic activity is known.[20] For benchmark paclitaxel and docetaxel are also included (labeled as TX3-A and TT1-A, respectively).

Our results showed that it is possible to directly correlate some molecular quantum descriptors to taxoids biological activity. This information can be used, in principle, to design new and more effective compounds.

METHODOLOGY

The taxoid compounds are formed from unusual structures with a phenylisoserine group esterifying an oxetane-containing diterpenoid at the position C-13. Paclitaxel and docetaxel (Figures 1 and 2) differ only by an acetate group attached to carbon C-10 of the moiety and a benzoate connected to nitrogen N(−C3′) substituent of the C-13 side chain.

We labeled the investigated structures (Figure 3, Table 1) as TX (paclitaxel or Taxol analogues)/TT (docetaxel or Taxotere analogues) considering the presence/absence of the previously mentioned acetate substituent at C-10 position. All the 20 studied compounds[20] are structurally very similar, some differing only by the S or R stereochemistry of the groups of chiral phenylisoserine carbons. These 20 molecules constitute a subset of the molecules investigated by Guéritte-Voegelein et al.[20] They were selected accordingly to the following criterion:

− to present a common “core” (taxane skeleton)

− to contain in the phenylisoserine side chain the groups (OH, $NHCO_2tBu$, NHCOPh) that are present in paclitaxel and docetaxel (basis of taxoid family).

The compounds can be divided into two groups of active and inactive molecules. We consider active (inactive) the molecules presenting antimitotic activity of ID_{50} <10 (ID_{50} $\geq$ 10). The biological in vitro index ID_{50} indicates the drug concentration (μM) necessary to decrease in 50% the microtubules depolymerization.

We have started our theoretical analysis carrying out fully geometrical optimizations for the 20 structures shown in Table 1. The number of compounds, their size, and conformational flexibility (large number of conformers with almost the same energy) make the geometry optimization a very expensive and difficult computational process, precluding the use of sophisticated ab initio methods.

The importance and necessity of carrying out detailed conformational searches for these kinds of compounds have been evidenced in a recent work[25] reporting a comparative study of semiempirical methods AM1 (Austin method one),[26] PM3 (Parametric method 3),[27–29] and DFT (density functional theory)[30] on the geometrical structure of 10-deacetylbaccatin-III, a taxoid precursor. Due to computational costs, the reported DFT calculations have used X-ray data[30] to generate tentative structural models. The AM1 and PM3 results have not only reproduced very well the DFT results but also indicated that the particular structure obtained by DFT had been in fact a local minima structure. These results clearly pointed out that the use of X-ray data in the generation of approximate structure to carry out geometrical optimizations could be misleading, and systematic conformational searches must be performed.

For the taxoids indicated in Table 1, with a larger number of atoms and more conformational flexibility than 10-deacetylbaccatin-III, the conformational search is mandatory. The geometrical stability of a set of conformations obtained by varying systematically the most important and flexible dihedral angles of the C-13 side chain was investigated in details. The minutiae of the conformational phase space analysis, including a comparison between the available experimental X-ray data[12,13] and the theoretical geometrical data for paclitaxel and docetaxel as well as the data for the electronic crystallographic density,[31] will be presented elsewhere.[32]

Although it is usually agreed that AM1 and PM3 are in general better than MNDO, there is an ongoing debate about the relative merits of AM1 and PM3. Some studies are in favor of PM3[33–35] and some of AM1.[36] Based on the comparative study reported for 10-deacetylbaccatin-III[25] we concluded that for the present study the PM3 is the best option.

All the PM3 calculations were carried out using the MOPAC 6.0 package[29] contained in the Chem2Pac[37] and Spartan[38] softwares. The optimized geometries were obtained setting the gradient in the energy hyper surface to be lower (in module) than 0.01 kcal/mol, to ensure good quality results.

Once the optimized geometries, eigenvalues, and eigenvectors had been obtained we proceeded to the density of states (DOS) calculations.

The electronic density of states (DOS) is defined as the number of electronic states per energy unit.[39] The related concept of local density of states (LDOS), i.e., the DOS calculated over a specific molecular region, is introduced in order to also describe the spatial distribution. The concept of LDOS allows us to map the distribution of any molecular orbital, occupied or virtual, over the molecular skeleton. LDOS gives an indication of how an atom or a set of them contributes to the formation of a specific molecular orbital. This can provide information on the contributions of specific molecular geometrical regions to the chemical reactivity, optical response, etc., and consequently, to their biochemical behavior. The LDOS approach has some advantages over conventional 3D rendering MO contours, such as lower computational cost, easier analysis, simultaneous determinations of relative contribution from various molecular levels, and results without resolution plot dependence.

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 out over the selected atomic orbitals (n_i to n_f), leading to the following expression to a specific energy level:[40–43]

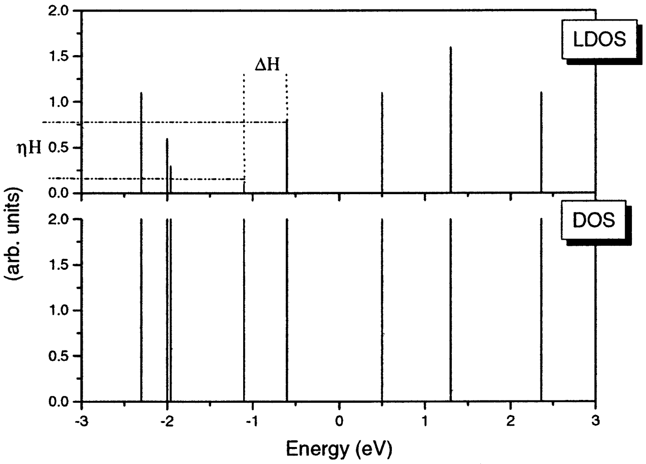

Figure 4. Typical LDOS and DOS discrete distributions. Each vertical line represents a state and its height indicates how much a selected region contributes to the molecular orbital associated with the energy states. ΔH represents the energy difference between two states and ηH the difference between their heights. See text for discussions.

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

The factor 2 comes from the Pauli exclusion principle (maximum of 2 electrons per electronic level). The c_{mi} came from expansion coefficients of molecular orbitals that in PM3 are expressed as a linear combination of atomic orbitals (LCAO approximation):[39]

$$\psi_m = \sum_i c_{mi}\varphi_i \tag{2}$$

This LDOS formulation is a discrete modulation, which allows a direct comparison of DOS and LDOS calculated from any LCAO method.

Figure 4 shows a typical DOS and LDOS results with each vertical line representing a state (discrete energy). For the DOS calculations the summation is over all the atoms of a molecule. For LDOS (in fact a projected DOS) the summation is carried out only for a selected group of atoms.

Once the DOS and LDOS had been calculated we have then carried out the structure–activity relationship analysis (SAR) using the Electronic Indices Methodology (EIM).[40]

The EIM has been employed with success in classificatory problems of different molecular systems, including carcinogens,[40–44] antitumoral and antibiotics,[45] steroid hormones,[46,47] and protease inhibitor[48] compounds.

The EIM main goal is to discriminate active and inactive compounds (with relation to some specific biological activity) using "universal" molecular quantum descriptors. EIM uses very simple Boolean rules based on the concept of local density of states and critical values for the electronic energy difference for the molecular orbital levels.

The EIM approach is based on two major descriptors, named η and Δ (Figure 4). The former, η, is related to the relative contributions difference between the most relevant molecular electronic levels over the identified molecular region linked to the biological activity (via LDOS):

$$\eta = 2\sum_{m=n_i}^{n_f} (|c_{m\mathrm{Level1}}|^2 - |c_{m\mathrm{Level2}}|^2) \qquad (3)$$

The criterion for selecting molecular regions for the LDOS analysis was based on the fact that these regions should be present in all taxoids and should be of special chemical significance (sites with attached groups, preferential sites for electrophilic/nucleophilic attacks, binding or docking regions, etc.).

The second major EIM descriptor is Δ. This descriptor is defined as the energy separation of the molecular levels indicated in eq 3:

$$\Delta = E_{\mathrm{Level1}} - E_{\mathrm{Level2}} \qquad (4)$$

Δ contains a global molecular information (involving the eigenvalues), while η contains local molecular information (LDOS), believed to reflect (on average) the molecular environmental biochemical mechanisms.

With only these two descriptors it is possible to derive simple rules to classify active and inactive compounds. Usually these rules have the following form:

IF $\eta > \eta_C$ AND $\Delta > \Delta_C$, the molecule will be active; otherwise it will be inactive.

η_C and Δ_C are critical values determined from EIM analysis. Based on the η and Δ parameters we can construct Boolean tables (relational conditions) for biological activity prediction. In contrast with other SAR methods EIM does not use training sets. The critical values for the Boolean rules are automatically obtained from an exploratory scanning.

EIM has been contrasted to more standard SAR methods such as Principal Component Analysis (PCA)[49] and Neural Networks (NN) methods[50] producing the same level of predictive accuracy but using fewer variables with significant reduction of computational efforts. In this work we have adopted a similar procedure contrasting the EIM results with the ones obtained from Principal Components Analysis (PCA)[49] and Hierarchical Cluster Analysis (HCA)[51] methods using the same framework of parameters.

PCA and HCA are methods widely used in pattern-recognition studies. PCA[52] is an extremely useful explorative tool, which maps samples through scores and individual descriptors by the loadings in a new vector space defined by the principal components (PC). In a geometrical interpretation, the set of descriptors of each molecule define its geometrical position (a molecule-point) in an n-dimensional space, where each descriptor corresponds to an axis. Score plots allow sample identification, clarifying whether they are similar or dissimilar, typical or outliers. From loading plots the most important descriptors can be easily identified as well as the correlation patterns among them. The first principal component is determined looking for the direction in the n-dimensional space which minimizes its distance to the molecule-points (direction of the maximum residual variance). From the remaining data variance (after the removal of the first PC) a second PC that is completely uncorrelated (orthogonal) with the first one is extracted and accounts for the maximum possible remaining data set variance. The procedure is then repeated until all PCs are generated. This corresponds to an n-dimensional space, isomorph to the n-dimensional variable space (n is the total number of used variables).

HCA,[51] also an exploratory tool, is used to validate the grouping previously identified by PCA. The primary goal of HCA is to emphasize the natural grouping of similar samples based on their proximity in the multidimensional space spanned by the used variables. The results, qualitative in nature, are presented in the form of a dendogram, allowing the visualization of clusters and correlation among samples. In HCA, the distances (Euclidean metric) between the samples are calculated and transformed into a similarity matrix whose elements are the similarity indexes ranging from zero to one; a smaller distance means a larger index.

The PCA and HCA studies were carried out using the program package einsight[53] where we have selected initially a set of 91 variables as potential descriptors to biological activity investigation.

RESULTS AND DISCUSSIONS

As mentioned above, the EIM main goal is to discriminate active and inactive compounds (with relation to some specific biological activity) using its two "universal" molecular quantum descriptors through very simple Boolean rules.

In the EIM approach it is essential the determination of the common regions of the molecular sets where the η descriptor will be evaluated. In the present work we have analyzed many regions including (see Figure 1):

- The oxetane ring (*oxet*).
- The core of the molecular skeleton composed by the A, B, and C rings (*core*).
- The benzoyl(−C2) substituent (*C2bez*).
- The Ph(−C3′) substituent of the side chain (*C3′ph*).
- The skeleton of the phenylisoserine chain composed by the atoms C1′, C2′, and C3′ and bonded oxygen atoms (*chain*).

The substituents NHO−Ph(−C3′) (present in the paclitaxel) and NHO−Tbut(−C3′) (present in the docetaxel) have not been considered as important regions because they are not elements of all molecules, one being absent when the other is present.

We have collected and analyzed the LDOS contribution for these five distinct regions, and the data obtained for the *chain* region gave us the best correlation with the biological indices. It is worth mentioning that it has been experimentally established that this side chain is structurally essential for active taxoids.

Also, after an exploratory search involving the frontier orbitals we determined that the HOMO and HOMO-1 are the best molecular levels to be used in the Δ parameter calculations.

Thus, the eqs 3 and 4 became

$$\eta H = 2\sum_{m=n_i}^{n_f} (|c_{m\mathrm{HOMO}}|^2 - |c_{m\mathrm{HOMO-1}}|^2) \qquad (5)$$

$$\Delta H = E_{\mathrm{HOMO}} - E_{\mathrm{HOMO-1}} \qquad (6)$$

where now m refers to the atomic orbitals associated with the *chain* region.

In Table 2 we show a summary with the main EIM data related to the Δ and η descriptors.

Table 2. EIM Electronic Variables: HOMO (H) and HOMO-1 (H-1) Energies and Their Difference (Δ) and the Contribution of the *Chain* Atoms to the H and H-1 Orbitals (C(H) and C(H_1)) and Their Difference (η)

molecules	H (eV)	H-1 (eV)	Δ (eV)	C(H) *chain*	C(H_1) *chain*	η *chain*
TX1-A	−9.7488	−9.9101	0.1613	0.0128	0	0.0128
TX1-B	−9.847	−10.1292	0.2822	0.0134	0.0396	−0.0262
TX2-A	−9.951	−10.0334	0.0824	0.0201	0.1632	−0.1431
TX2-B	−9.8967	−9.9292	0.0325	0.1206	0.1077	0.0129
TX3-A	−9.7372	−9.951	0.2138	0.0155	0	0.0155
TX3-B	−9.7269	−9.8694	0.1425	0.0081	0.0792	−0.0711
TX4-A	−9.8676	−9.9224	0.0548	0.1879	0.0234	0.1645
TX4-B	−9.8707	−9.9135	0.0428	0.1991	0.0519	0.1472
TT1-A	−9.722	−9.9438	0.2218	0.0125	0	0.0125
TT1-B	−9.8436	−10.13	0.2864	0.0135	0.0373	−0.0238
TT1-C	−9.9937	−9.998	0.0043	0.0136	0.0047	0.0089
TT1-D	−9.678	−10.0068	0.3288	0.0134	0	0.0134
TT2-A	−9.94	−10.0333	0.0933	0.0197	0.1587	−0.139
TT2-B	−9.7336	−9.9036	0.17	0.0183	0.2539	−0.2356
TT3-A	−9.7322	−9.9944	0.2622	0.0152	0	0.0152
TT3-B	−9.7497	−9.8885	0.1388	0.0082	0.0764	−0.0682
TT3-C	−9.8713	−9.8865	0.0152	0.0046	0.0163	−0.0117
TT3-D	−9.7503	−9.9578	0.2075	0.0145	0.0124	0.0021
TT4-A	−9.868	−9.9181	0.0501	0.1799	0.0269	0.153
TT4-B	−9.874	−9.9119	0.0379	0.165	0.086	0.079

Table 3. Table of Taxoids Classification According to the EIM Rules[a]

molecule	Δ	η	EIM	Q.A.
TX1-A	+	+	A	A
TX1-B	+	-	I	I
TX2-A	-	-	*A*	*I*
TX2-B	-	+	I	I
TX3-A	+	+	A	A
TX3-B	-	-	A	A
TX4-A	-	+	I	I
TX4-B	-	+	I	I
TT1-A	+	+	A	A
TT1-B	+	-	I	I
TT1-C	-	+	*I*	*A*
TT1-D	+	+	A	A
TT2-A	-	-	*A*	*I*
TT2-B	+	-	I	I
TT3-A	+	+	A	A
TT3-B	-	-	A	A
TT3-C	-	-	A	A
TT3-D	+	+	A	A
TT4-A	-	+	I	I
TT4-B	-	+	I	I

[a] The use of ± symbols follows the critical values and Boolean rules defined below. The italicized entries show the molecules to which the EIM and experimental classification are in disagreement.

From this table we can easily identify the critical values and derive simple relation Boolean rules to classify active and inactive taxoids:

Critical Values:

− $\Delta > 0.15$ (+); $\Delta \leq 0.15$ (−);

− $\eta > 0.0$ (+); $\eta \leq 0.0$ (−);

Boolean Rules:

Δ(+) and η (+) ⇒ (+) active

Δ (+) and η (−) ⇒ (−) inactive

Δ (−) and η (+) ⇒ (−) inactive

Δ (−) and η (−) ⇒ (+) active

+/− means that the Δ and η values are greater/lower than the critical values.

The Boolean rules have been used to compare the theoretical predictions with the available experimental data. These results are summarized in Table 3. Seventeen out of 20 molecules have been correctly classified (85% accuracy). Only one active and two inactive compounds have been incorrectly classified.

It is very impressive that using only two electronic parameters from *in a vacuum* molecular calculations we could obtain a very significant correlation with the taxoid biological activity, which is a very complex biochemical phenomenon.

We would like to stress that the same order of predictive accuracy have been obtained with EIM (using the same descriptors) for thousands of organic compounds presenting diversified biological activity such as carcinogens,[40−44] antibiotics and antitumoral,[45] protease inhibitors,[48] hormones,[46,47] etc.

The same level of predictive accuracy observed for other and diversified classes of organic compounds[40−48] and the great similarity between the rules used to classify active and inactive compounds proved the reliability and large applicability of EIM approach corroborating the “universal” character of their molecular quantum descriptors. This validates the EIM approach as a new, low cost, and very effective SAR methodology. It is worth mentioning that although the EIM analysis is low cost and very fast, obtaining the quantum descriptors (which is not dependent on the SAR approach) might be, as in the present case, a costly computational process.

However, it remains to be elucidated why these descriptors are so important. Although we can correlate the η parameter as an indirect measure of the local net charge, and consequently chemical reactivity, no specific chemical or physical property can be associated with the Δ parameter. Preliminary calculations for PAHs indicated that this descriptor could perhaps be understood as an indirect measure of the chemical reactivity and lifetime of excited states in the process of charge-transfer and/or covalent bond formations to PAH−DNA adducts. However these arguments do not apply to other classes or organic classes where Δ is also a major descriptor allowing distinguish active and inactive compounds. They have only in common the fact that Δ is expressed in terms of the Schrödinger’s equation eigenvectors. Perhaps this Δ “universality” class might be related to the existence of common topological features (fractal/

chaotic/scaling).[54] Further studies are needed to clarify these aspects.

To have an independent statistical validation for the EIM results and electronic descriptors we have carried out PCA and HCA analysis for the same molecular set.

We have started from an initial set of 91 potential relevant descriptors. This set involved the EIM descriptors and other physicochemical parameters such as the following:

− Energies of the frontier molecular orbitals, their next levels, and their energy differences.

− Local density of states (LDOS) over the selected molecular regions mentioned above and their differences.

− Hardness, approximated as HD = (LUMO−HOMO)/2.

− Mulliken electronegativity, approximated as $\chi = -(\mathrm{HOMO}+\mathrm{LUMO})/2$.

− Dipole moment.

− Refractivity.

− Polarizability.

− Mass.

− Hydration energy.

− Coefficient of molecular partition octanol−water (log P).

− Solvatation energies.

Parameters such refractivity, polarizability, mass, and the log P have been disregarded after a preliminary analysis, due in part to their inability to differentiate isomers presenting distinct biological activity.

PCA/HCA analyses have produced various sets of molecular descriptors presenting 100% of accuracy in separating active and inactive compounds. In general, descriptors such as the listed below were present:

− C(H-1) *chain*, C(H) *chain*, C(H-1) *C3′Ph*, C(H-2) *oxet*, representing the LDOS of frontier orbitals for specific regions,

− ηH *chain*, ηH′*chain*, ηH″*chain*, ηH′*C2bez*, ηL″*C3′Ph*, representing the difference between LDOS of the frontier orbitals for specific regions

− ΔH, representing the energy difference HOMO − (HOMO-1), where HOMO refers to Highest Occupied Molecular Orbital.

− ΔL, representing the energy difference (LUMO+1) − LUMO, where LUMO refers to Lowest Unoccupied Molecular Orbital.

These parameters were mainly related to the EIM descriptors. When the EIM descriptors were not used, the PCA predictive power was significantly reduced. These results provided further evidence about the EIM statistical value and added new data on the class of "universality" of the EIM descriptors.

In Figure 5 we show the score of the first two principal components (PC1 and PC2) for the 20 taxoids for one of the typical obtained PCA analysis. The descriptors have been auto scaled prior to the analysis. The molecules have been separated into two subsets (active and inactive), vertically opposed, indicating that PC2 was the major responsible for this characterization. The active group is located on the bottom of the plot and the inactive one on the top.

The PC1 and PC2 axes corresponded to 51.38 and 26.56% of the system variance, respectively. The plane PC1xPC2 conserved 77.94% of the total variance of the original data, and the axis were written in terms of molecular descriptors according to the following equations:

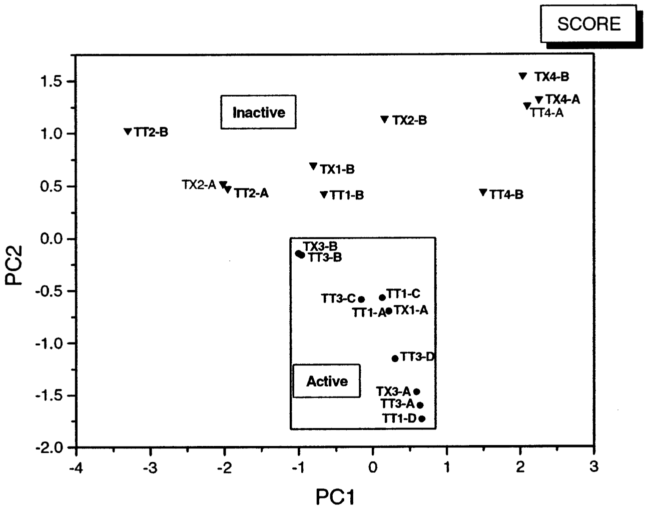

Figure 5. Plot of two PC vectors to taxoids score. The 20 compounds are separated into two subgroups as active and inactive. All molecules were correctly classified.

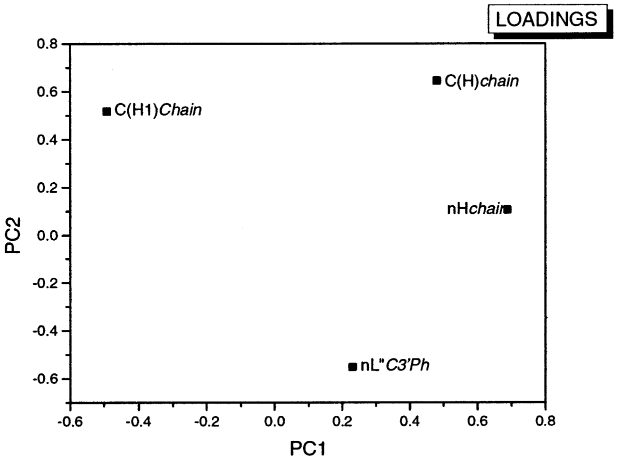

Figure 6. Plot of two PC vectors to loading descriptors. The four descriptors related to the taxoids classification as active or inactive are presented.

$$\mathrm{PC1} = 0.688\ \eta\mathrm{H}\ \mathit{chain} + 0.481\mathrm{C(H)}\ \mathit{chain} + 0.231\ \eta\mathrm{L}''\ \mathit{C3'Ph} - 0.491\ \mathrm{C(H\text{-}1)}\ \mathit{chain} \quad (7)$$

$$\mathrm{PC2} = 0.645\ \mathrm{C(H)}\ \mathit{chain} + 0.517\ \mathrm{C(H\text{-}1)}\ \mathit{chain} + 0.104\ \eta\mathrm{H}\ \mathit{chain} - 0.553\ \eta\mathrm{L}''\ \mathit{C3'Ph} \quad (8)$$

It is interesting to notice that the highest contribution to the PC1 and PC2 has been from the LDOS of the occupied orbitals over *chain* atoms, exactly a variable of the EIM analysis.

Figure 6 shows the loading vectors of the descriptors that have oriented the molecules separation. The plot indicates that the descriptors C(H) *chain* and ηL″*C3′Ph* have been the principal responsibles for pulling the active and inactive molecules toward to distinct regions of the graphic.

From an analysis combining the data in Figure 6 and eq 8 we can see that more active molecules can be obtained when we have small (positive) values for the variables C(H) *chain*, C(H-1) *chain*, combined with negative values for the variable ηH *chain* and positive values to ηL″*C3′Ph*.

Corroborating the PCA results we present in Figure 7 the HCA plot. The theoretical descriptors have been auto scaled,

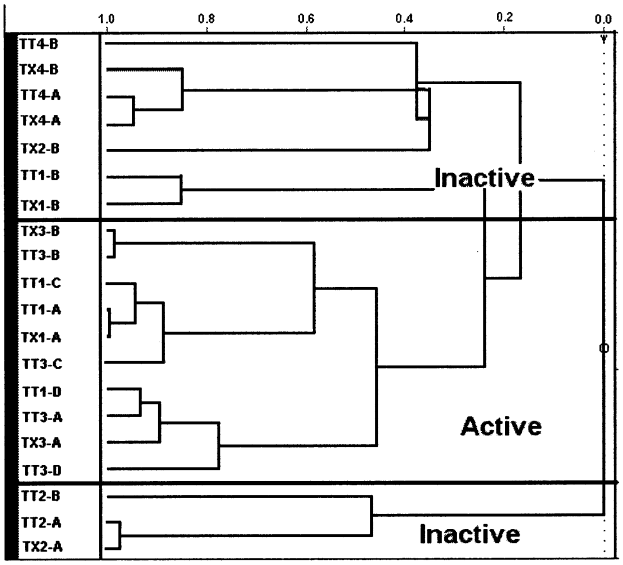

Figure 7. Dendogram of the hierarchical distribution of the taxoids. The active and inactive compounds are separated being the former grouped

and the molecules have been grouped considering the centroid Euclidian distance. The horizontal lines represent the compounds, while the vertical ones show the similarity between pairs of them. The active molecules have been grouped in the middle of this allocation (beginning with the TX3-B and ending with the TT3-D). This very close distribution shows that the adopted descriptors have been able to sustain the necessary similarity for all the active taxoids. The inactive molecules, on the other hand, were not all grouped and had a very small similarity (0.24) with the actives.

These results, selecting the four variables correlated with the activity of these 20 taxoids, can be used to assist the design of new molecules of this class.

These two pattern recognition methods (PCA and HCA) have classified the 20 compounds into two groups exactly as EIM had done and consistently identified the most relevant descriptors.

In summary we have presented a study of the EIM and PCA/HCA methodologies able to correctly classify active and inactive taxoids with 100% of accuracy using only a few “universal” quantum molecular descriptors. It was possible to identify the necessary electronic features defining active molecules. This information can be used to select and design new active compounds.

We can use the EIM in a “combinatorial” sense. This is, we can vary the groups and through the EIM patterns identify potential active/inactive compounds. We can also use the information gained from the EIM parameters to carry out more specific chemical substitutions that would preserve/increase the features required to a specific molecule matches the pattern of active ones. The combined use of EIM with PCA/HCA can be a new and very efficient tool in the field of computer assisted drug design.

REFERENCES AND NOTES

(1) *The Biological Basis of Cancer*; McKinnel, R. G., Ed.; Cambridge University Press: Cambridge, 1998.

(2) Lodish, H.; Baltimore, D.; Berk, A.; Zipursky, S. L.; Sudaira, P. M.; Darnell, J.; *Molecular Cell Biology*; Scientific American Books: New York, 1995.

(3) *Cancer Biology*; Ruddon, R. W., Ed.; Oxford University Press: New York, 1987.

(4) Sugimura, T. Multistep carcinogenesis: A 1992 perspective. *Science* **1992**, *258*, 603−607.

(5) EIU Marketing in Europe, Trade Reviews, 337, December 1990.

(6) *Taxane Anticancer Agents.* Georg, G. I., Chen, T. T., Ojima, I., Vyas, D. M., Eds.; ACS Symposium Series 583; American Chemical Society: Washington, DC, 1995.

(7) Wani, M. C.; Taylor, H. L.; Wall, M. E.; Coggon, P.; McPhail, A. T. Plant antitumor agents .6. Isolation and structure of taxol, a novel antileukemic and antitumor agent from Taxus-Brevifolia. *J. Am. Chem. Soc.* **1971**, *93*, 2325−2327.

(8) Guénard, D.; Guéritte-Voegelein, F.; Potier, P. Taxol and Taxotere − Discovery, chemistry, and structure−activity-relationships. *Acc. Chem. Res.* **1993**, *26*, 160−167.

(9) Commerçon, A.; Bouzart, J. D.; Didier, E.; Lavelle, F. In *Taxane Anticancer Agents;* Georg, G. I., Chen, T. T., Ojima, I., Vyas, D. M., Eds.; ACS Symposium Series 583; American Chemical Society: Washington, DC, 1995; Chapter 17, pp 233−246.

(10) Nicolaou, K. C.; Yang, Z.; Liu J. J.; Ueno, H.; Nantermet, P. G.; Guy, R. K. et al. Total synthesis of Taxol. *Nature* **1994**, *367*, 630−634.

(11) Altmann, K.-H.; Wartmann, M.; O’Reilly, T. Epothilones and related structures -a new class of mocrotubules inhibitors with potent in vivo antitumor activity. *Biochim. Biophys. Acta* **2000**, *1470*, M79−M91.

(12) Guérrite-Voegelein, F.; Mangatal, L.; Guénard, D.; Potier, P.; Gilhem, J.; Cesario, M.; Pascard, C. Structure of synthetic taxol precursor: N-*tert*-butoxycarbonyl-10-deacetyl-N-debenzoyltaxol. *Acta Crystallogr.* **1990**, *C46*, 781−784.

(13) Mastropaolo, D.; Camerman, A.; Luo, Y.; Brayer, G. D.; Camerman, N. Crystal and molecular structure of paclitaxel (taxol). *Proc. Natl. Acad. Sci. U.S.A.* **1995**, *92*, 6920−6924.

(14) Gao, Q.; Parker, W. L. The “hydrophobic collapse”conformations of paclitaxel (Taxol) has been observed in a nonaqueous environment: crystal structure of 10-deacetyl-7-epitaxol. *Tetrahedron* **1996**, *52*, 2291−2300.

(15) Snyder, J. P.; Nevins, N.; Cicero, D. O.; Jansen, J. The conformation of taxol in chloroform. *J. Am. Chem. Soc.* **2000**, *122*, 724−725.

(16) Williams, H.; Moyna, G.; Scott, A. I. NMR and molecular modeling study of the conformations of taxol 2′-acetate in chloroform and aqueous dimethyl sulfoxide solutions. *J. Med. Chem.* **1996**, *39*, 1555−1559.

(17) Falzone, C. J.; Benesi, A. J.; Lecomte, J. T. Characterization of taxol in methylene chloride by NMR spectroscopy. *Tetrahedron Lett.* **1992**, *33*, 1169−1172.

(18) William, H. J.; Scott, A. I.; Dieden, R. A. NMR and molecular modeling study of the conformations of taxol and of its side chain methylester in aqueous and nonaqueous solution. *Tetrahedron* **1993**, *49*, 6545−6560.

(19) Zhu, Q.; Guo, Z.; Huang, N.; Wang, M.; Chu, F. Comparative molecular field analysis of a series of paclitaxel analogues. *J. Med. Chem.* **1997**, *40*, 4319−4328.

(20) Guéritte-Voegelein, F.; Guénard, D.; Lavelle, F.; Le Goff, M.-T.; Mangatal, L.; Potier, P. Relationships between the structure of taxol analogues and their antimitotic activity. *J. Med. Chem.* **1991**, *34*, 992−998.

(21) Kingston, D. G. I.; Chaudhary, A. G.; Chordia, M. D.; Gharpure, M.; Gunatilaka, A. A. L. Synthesis and biological evaluation of 2-acyl analogues of paclitaxel (taxol). *J. Med. Chem.* **1998**, *41*, 3715−3726.

(22) Ojima, I.; Slater, J. C.; Michaud, E.; Kuduk, S. D.; Bounaud, P.-Y.; Vrignaud, P.; Bissery, M.-C.; Veith, J. M.; Pera, P.; Bernacki, R. J. Synthesis and structure−activity relationships of the second-generation antitumor taxoids: exceptional activity against drug-resistant cancer cells. *J. Med. Chem.* **1996**, *39*, 3889−3896.

(23) Ojima, I.; Bounaud, P.-Y.; Ahern, D. G. New photoaffinity analogues of paclitaxel. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 1189−1194.

(24) Roh, E. J.; Kim, D.; Choi, J. Y.; Lee, B.-S.; Lee, C. O.; Song, C. E. Synthesis, biological activity and receptor-based 3-D QSAR study of 3′-N-substituted-3′-N-debenzoylpaclitaxel analogues. *Bioorg. Med. Chem.* **2002**, *10*, 3135−3143.

(25) Braga, S. F.; Galvão, D. S. A semiempirical study on the electronic structure of 10-deacetylbaccatin-III. *J. Molecular Graphics Modeling* **2002**, *21*, 57−70.

(26) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The development and use of quantum -mechanical molecular -models. 76. AM1 − A new general -purpose quantum -mechanical molecular -model. *J. Am. Chem. Soc.* **1985**, *107*, 3902.

(27) Stewart, J. J. P.; Optimization of parameters for semiempirical methods. 1. Method. *J. Comput. Chem.* **1989**, *10*, 209−220.

(28) Stewart, J. J. P.; Optimization of parameters for semiempirical methods. 2. Applications. *J. Comput. Chem.* **1989**, *10*, 221−264.

(29) MOPAC Program, version 6.0, Quantum Chemistry Program Exchange No. 455. http://qcpe.chem.indiana.edu

(30) Ballone, P.; Marchi, M. A density functional study of a new family of anticancer drugs: paclitaxel, taxotere, ephotilone and discodermolide. *J. Phys. Chem. A* **1999**, *103*, 3097−3102.

(31) Snyder, J. P.; Nettles, J. H.; Cornett, B.; Downing, K. H.; Nogales, E. The binding conformation of taxol in β-tubulin: a model based on electron crystallographic density. *Proc. Natl. Acad. Sci. U.S.A.* **2001**, *98*, 5312−5316.

(32) Braga, S. F.; Galvão, D. S. to be published.

(33) Budyka, M. F., Zyubina, T. S.; Ryabenko, A. G. Computer modeling of C-2 cluster addition to fullerene C-60. *Intl. J. Quantum Chem.* **2002**, *88*, 652−662.

(34) Lozynski M.; Rusinska-Ruszak, D. PM3 conformations of C-13 Taxol side chain methyl ester. *Tetrahedron Lett.* **1995**, *36*, 8849−8852.

(35) Baumann, H.; Martin, R. E.; Diederich, F. PM3 geometry optimization and CNDO/S-CI computation on UV/VIS spectra of large organic structures: program description and application to poly(tracetylene) hexamer and taxotere. *J. Comput. Chem*. **1999**, *20*, 396−411.

(36) Dávila, L. Y. A.; Caldas, M. J. Applicability of MNDO techniques AM1 and PM3 to ring-structured polymers. *J. Comput. Chem.* **2002**, *23*, 1135−1142.

(37) Cyrillo, M.; Galvão, D. S. Chem2Pac: A computational chemistry integrator for Windows. *EPA Newsletter* **1999**, *67*, 31−34 and 34−38. http://www.ifi.unicamp.br/gsonm/chem2pac.

(38) Pc Spartan Pro 1.0.3. 2000 Wavefunction, Inc. http://www.wavefun.com.

(39) Levine, I. N. *Quantum Chemistry*, 4th ed.; Prentice-Hall: Englewood Cliffs: NJ, 1991.

(40) Barone, P. M. V. B.; Camilo, A.; Galvão, D. S. Theoretical approach to identify carcinogenic activity of polycyclic aromatic hydrocarbons. *Phys. Rev. Lett.* **1996**, *77*, 1186−1189.

(41) Barone, P. M. V. B.; Braga, R. S.; Camilo A.; Galvão, D. S. Electronic indices from semiempirical calculations to identify carcinogenic activity of polycyclic aromatic hydrocarbons. *J. Mol. Struct*. (THEOCHEM) **2000**, *505*, 55−66.

(42) Vendrame, R.; Braga, R. S.; Takahata, Y.; Galvão, D. S. Structure−activity relationship studies of carcinogenic activity of polycyclic aromatic hydrocarbons using calculated molecular descriptors with principal component analysis and neural network methods. *J. Chem. Inf. Comput. Sci*. **1999**, *39*, 1094−1104.

(43) Vendrame, R.; Braga, R. S.; Takahata, Y.; Galvão, D. S. Structure-carcinogenic activity relationship studies of polycyclic aromatic hydrocarbons (PAHs) with pattern-recognition methods. *J. Mol. Struct.* (THEOCHEM) **2001**, *539*, 253−265.

(44) Coluci, V. R.; Vendrame, R.; Braga, R. S.; Galvão, D. S. Identifying Relevant Molecular Descriptors Related to Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons (PAHs) Using Pattern Recognition Methods *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1479−1489.

(45) Santo, L. L. E.; Galvão, D. S. Structure−activity study of indolequinones bioreductive alkylating agents. *J. Mol. Struct*. (THEOCHEM) **1999**, *464*, 273−279.

(46) Vendrame, R.; Coluci, V. R.; Braga, R. S.; Galvão, D. S. Structure−activity relationship (SAR) studies of the tripos benchmark steroids. *J. Mol. Struct. (THEOCHEM)* **2002**, *619*, 195−205.

(47) Braga, R. S.; Vendrame, R.; Galvão, D. S. Structure−activity relationship studies of substituted 17 alpha-acetoxyprogesterone hormones. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 1377−1385.

(48) Cyrillo, M.; Galvão, D. S. Structure−activity relationship study of some inhibitors of HIV-1 integrase. *J. Mol. Struct*. *(THEOCHEM)* **1999**, *464*, 267−272.

(49) Naes, T.; Baardseth, P.; Helgesen, H.; Isakson, T. Multivariate techniques in the analysis of meat quality. *Meat Sci*. **1996**, *43*, s135−s149.

(50) Hagan, M. T.; Demuth, H. B.; Beale, M. *Neural Network Design*; PWS Publishing Company: Boston, 1996.

(51) Beebe, K. R.; Pell, R. J.; Seasholtz, M. B. *Chemometrics − A Practical Guide*; Wiley: New York, 1998.

(52) Massart, D. L.; Vandeginste, B. G. M.; Deming, S. N.; Michotte, Y.; Kaufman, L. In *Chemometrics: a textbook 2*. Elsevier: Chapter 21, p 369

(53) Einsight 3.0. Infometrix, Inc. 2200 Sixth Ave, Suite 833, Seattle WA 98121, 1991.

(54) Dewey, T. G. *Fractals in Molecular Biophysics*; Oxford University Press: Oxford, 1997.

CI025640V