J Mol Model (2003) 9:395–407
DOI 10.1007/s00894-003-0148-7

ORIGINAL PAPER

Humberto Gonzáles-Díaz · Ornella Gia · Eugenio Uriarte · Ivan Hernádez · Ronal Ramos · Mayrelis Chaviano · Santiago Seijo · Juan A. Castillo · Lázaro Morales · Lourdes Santana · Delali Akpaloo · Enrique Molina · Maikel Cruz · Luis A. Torres · Miguel A. Cabrera

Markovian chemicals "in silico" design (MARCH-INSIDE), a promising approach for computer-aided molecular design I: discovery of anticancer compounds

Received: 20 March 2003 / Accepted: 7 July 2003 / Published online: 16 September 2003

Abstract A simple stochastic approach, designed to model the movement of electrons throughout chemical bonds, is introduced. This model makes use of a Markov matrix to codify useful structural information in QSAR. The self-return probabilities of this matrix throughout time ($^{SR}\pi_k$) are then used as molecular descriptors. Firstly, a calculation of $^{SR}\pi_k$ is made for a large series of anticancer and non-anticancer chemicals. Then, *k*-Means Cluster Analysis allows us to split the data series into clusters and ensure a representative design of training and predicting series. Next, we develop a classification function through Linear Discriminant Analysis (LDA). This QSAR discriminates between anticancer compounds and non-active compounds with a correct global classification of 90.5% in the training series. The model also correctly classified 86.07% of the compounds in the predicting series. This classification function is then used to perform a virtual screening of a combinatorial library of coumarins. In this connection, the biological assay of some furocoumarins, selected by virtual screening using the present model, gives good results. In particular, a tetracyclic derivative of 5-methoxypsoralen (5-MOP) has an IC_{50} against HL-60 tumoral line around 6 to 10 times lower than those for 8-MOP and 5-MOP (reference drugs), respectively. Finally, application of Iso-contribution Zone Analysis (IZA) provides structural interpretation of the biological activity predicted with this QSAR.

Keywords Markov chain · Molecular design · QSAR · Anticancer compounds · Linear discriminant analysis · Cluster analysis · Random process

H. Gonzáles-Díaz (✉) · I. Hernádez · M. Chaviano · S. Seijo · J. A. Castillo · L. Morales · D. Akpaloo · E. Molina · M. Cruz · L. A. Torres · M. A. Cabrera
Chemical Bioactives Center,
Central University of Las Villas,
54830 Santa Clara, Villa Clara, Cuba
e-mail: humbertogd@vodafone.es or humbertogd@cbq.uclv.edu.cu
Tel.: +53/42/281473-131
Fax: +53/42/281473-455

O. Gia
Department of Pharmaceutical Sciences,
University of Padua,
Via Marzolo 5, 35131 Padua, Italy

H. Gonzáles-Díaz · E. Uriarte · L. Santana
Department of Organic Chemistry, Faculty of Pharmacy,
University of Santiago de Compostela,
15782 Santiago de Compostela, Spain

R. Ramos
Department of Chemistry, Chemistry and Pharmacy Faculty,
Central University of Las Villas,
54830 Santa Clara, Villa Clara, Cuba

L. A. Torres
Department of Pharmacy, Chemistry and Pharmacy Faculty,
Central University of Las Villas,
54830 Santa Clara, Villa Clara, Cuba

Introduction

The use of so-called Markov's chains began at the beginning of the last century (1901). [1] Since this earlier work after Markov and up to 1960, different applications of the stochastic process in various fields of science appeared, including astronomy, physics, biology, and chemistry. [2] From the 1960s until today, there has been no decline in this explosion in the use of Markov's process. On the contrary, a continuous increase in the use of Markov's chains (MCH) theory is expected in the near future. [3] Some branches of science such as artificial intelligence, [4] epidemiology, [5] and medicine [6] have incorporated useful methods based on this mathematical approach.

In biological sciences, particularly bioinformatics and related subjects, the MCH models have proved to be largely useful. Markov models are well-known tools for analyzing biological sequence data and have been used in detecting new genes from open reading frames. [7, 8] Other uses of these models have included data based searching and multiple sequence alignment of protein families and protein domains. [9] Protein subcellular locations have been also successfully predicted. [10, 11] Hubbard and Park used amino acid sequence-based hidden Markov models for predicting secondary protein structures. [12] In this sense, Krogh et al. [13] also proposed their hidden Markov model architecture. Markov's stochastic process has also been used for protein folding recognition. [14]

Throughout time, the use of MCH has grown as rapidly as the particle cascades that they can describe. For example, MCH are used in quantum mechanics to resolve the many-electron problem by quantum Monte Carlo methods. [15] In any case, stochastic processes and matrices have been present in the foundations of quantum mechanics from the outset. In 1925, W. Heinsenberg introduced a quantum system representation, which later prompted the development of matrix mechanics by M. Born, W. Heinsenberg and P. Jordan. This representation describes the transition of the quantum system of particles (e.g. electrons) from one state to the other using transition frequencies or probabilities. [16] The probabilistic interpretation of quantum phenomena is a well-established point of view, also used in the Schrödinger representation [16] and density functional theory. [17]

The pharmaceutical industry is also under increasing pressure to discover new drugs, leading to faster and more efficient methods than those used in the past. In this case, molecular modeling and molecular structure codification techniques have emerged as a promising solution to this problem. [18, 19, 20, 21] This is the reason why different molecular descriptors have continuously appeared in the literature, including topological, informational, graph-theoretical, quantum mechanical, molecular mechanics-based molecular descriptors, amongst others. [18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33] A recently published handbook by Todeschini and Consonni offered a summary of many of them. [34]

In any case, there is still great interest in the development of new molecular descriptors. The broad diversity of chemical structures and biological activities that it is necessary to correlate by QSAR methods has been the driving force behind this interest. In particular, the search of anticancer compounds has always been on the desktop of molecular modeling and drug design specialists. In spite of this intensive search, the discovery of selective antitumor compounds has remained a largely elusive goal of cancer research. Subsequently, new approaches are needed in order to make an efficient search for candidates to be assayed as anticancer drugs. [35, 36, 37, 38, 39, 40, 41, 42, 43, 44]

However, when chemists try to apply quantum mechanics calculations to codify useful structural information in pharmacological terms, time becomes a limiting factor. As a result, many simple molecular descriptors are used to represent molecular structure. The simplicity of Markov chains as well as their stochastic nature therefore attracted our attention as a possible source of simple but physically meaningful molecular descriptors. As molecular descriptors, the authors of this paper understand simple numerical indices that are used to codify the molecular structure in Quantitative Structure Activity (Property) Relationship (QSAR and QSPR) studies. [45] In a recent paper, some authors of the present paper have additionally enlarged the limits of applicability of those molecular descriptors in QSAR. [46] These new approaches generally loose theoretical rigor (with respect to physical theories) but gain practical applicability, one of the starting points in the development of almost every novel molecular descriptor. [47, 48]

Nevertheless, the use of stochastic matrix formalism as a source of simple molecular descriptors did not appear in the literature before 2002. Last year, González et al. used a Markov chain formalism for the first time to codify molecular structure towards virtual screening, and rational experimental discovery of fluckicidal drugs. [49] These ideas have been extended to the study of protein structure property relationships. [50] Recent work reported the generalization of our molecular descriptors to codify 3D molecular structure without any loss of theoretical meaning. [51] Therefore, considering all of the issues highlighted in this introductory section, the present paper has very specific aims. Essentially, this paper deals with the QSAR study of anticancer activity of large and heterogeneous series of organic compounds in order to continue the validation of ${}^{SR}\pi_k$ as useful molecular descriptors. Secondly, the paper intends to apply the present QSAR for a virtual mining of a combinatorial library of coumarins to detect more active leads of this family of compounds. Consequently, those chemicals predicted with the highest activity will be re-synthesized and experimentally assayed. Finally, local calculations of the molecular descriptors will permit structural interpretation of the model when applying a simple method we called the Iso-contribution Zone Analysis (IZA). [46, 52]

Materials and methods

Markovian chemicals "in silico" design (MARCH-INSIDE)

The description we offer in this section constitutes the theoretical background of a simple but still physically meaningful and highly flexible model of intramolecular electron delocalization. The model explicitly codifies molecular connectivity and, at the same time, the effect of the presence of heteroatoms in electron distribution throughout the drug backbone. Both aspects appear to be very important features in QSAR. [53, 54, 55]

Consider a hypothetical situation in which a series of atoms are free in space at an arbitrary initial time (t_0). Alternatively, one may imagine a more real situation in which, after perturbation by some external factor, the electrons reach a distribution around atom cores different to that which they possess in the stationary state. It may

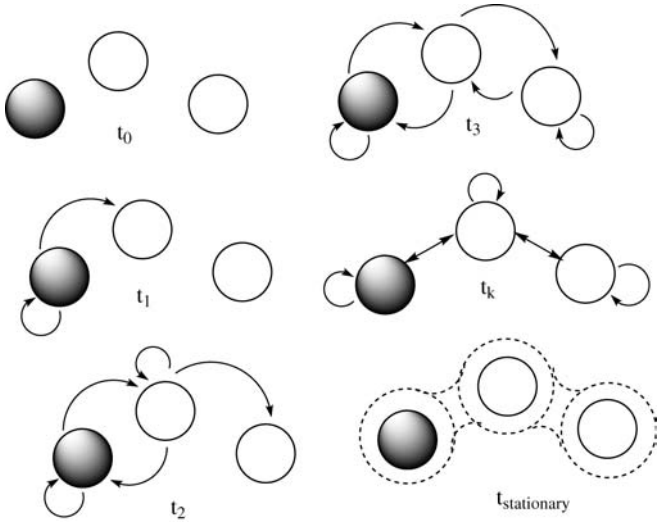

Fig. 1 Diagrammatic representation of random electron distribution in a simple Markovian model. The symbol $t_{stationary}$ represent the stationary time: the time at which electrons reach equilibrium distribution around atoms

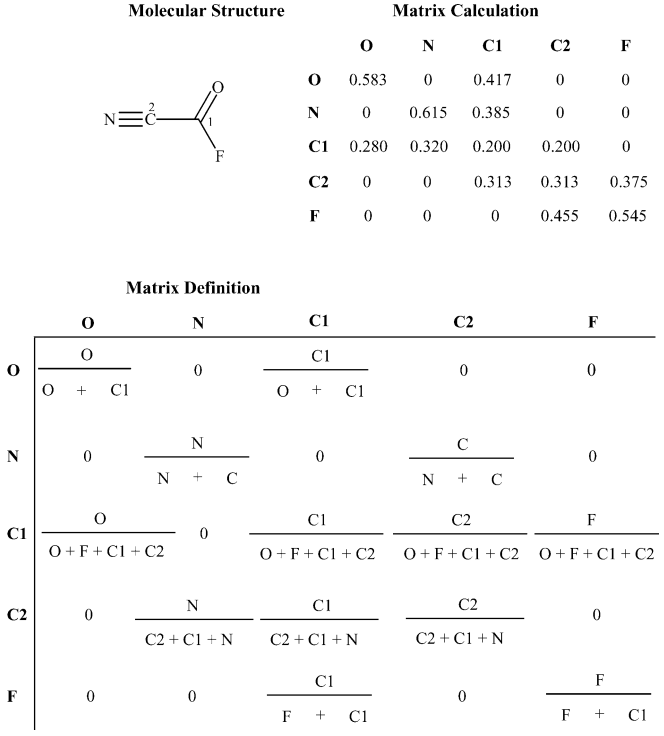

Fig. 2 Definition and calculation of the ${}^{1}\prod$ matrix for a specific case. The element symbol is used to denote the value of the element electronegativity, so for example: F=fluorine electronegativity χ(F)

therefore be interesting to develop a simple stochastic model of the return of electrons to the original position throughout time. A model of this type, closely related to molecular structure information, could act as a source of novel physically meaningful molecular descriptors.

Assume that after either of these initial situations, electrons start to distribute around atom cores in discrete intervals of time t_k. By using MCH [1, 2, 3, 49, 50, 51, 56] it is therefore possible to develop a simple model of the probabilities with which electrons move around these atom cores in further intervals of time, until a stationary electron density distribution appears (see Fig. 1). As depicted in Fig. 1, this model will describe the probabilities ($^{k}p_{ij}$) with which electrons move from any arbitrary atom a_i at time t_0 (in black) to other a_j atoms (in white) throughout discrete time periods t_k (k=1, 2, 3, ...) and throughout the chemical bonds. This model is stochastic per se (probabilistic distribution of electrons in time) but, as mentioned above, actually considers molecular connectivity (the distribution of electrons in space throughout the chemical bonds).

The selection of a Markov chain process is not arbitrary. From quantum physics, it is well known that, if electrons are labeled at an arbitrary initial time, one cannot use these labels to distinguish between them in subsequent moments. This physical fact has been historically referred to as the principle of the indistinguishability of identical particles. [16] An MCH-based model of electron distribution around atom cores obeys this principle perfectly, as one of the main characteristics of MCH is that the probability of occurrence of an event (electron movement) does not depend on the previous states of the system (the former atoms from which electrons came). [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 49, 50, 51, 56] This means that the model does not depend on any electron labeling.

The present procedure considers the external electron layers of any atom core in the molecule (valence shell) as states of the MCH. [49, 50, 51] The method uses the matrix ${}^{1}\Pi$, which has the elements $^{1}p_{ij}$. This matrix is called the 1-step electron-transition stochastic matrix. ${}^{1}\Pi$ is built as a square table of order n, where n represents the number of atoms in the molecule. The elements ($^{1}p_{ij}$) of the 1-step electron-transition stochastic matrix are the transition probabilities with which electrons move from atom i to j in the interval t_1=1 (considering t_0=0). The main simplification here, which may appear to be a drawback but is actually an advantage, is to suppose that electronegativity quantifies the strength with which the atoms restore the electrons to their stationary position:

$$ {}^{1}p_{ij} = \frac{{}^{I}\chi_j}{\sum_{k=1}^{\delta+1} {}^{I}\chi_k} \tag{1} $$

where ${}^{I}\chi_j$ is Pauling's electronegativity of the atom a_j, which is bonded to the atom a_i. [49, 50, 51, 57] The elements of ${}^{1}\Pi$ ($^{1}p_{ij}$) are defined to codify information about the electron-withdrawing strength of atoms to withdraw electrons from their neighbors in the molecule. We will only use ${}^{1}\Pi$ afterwards. Conversely, the p_{ij} values are inversely related to the electronegativity of the atoms that "compete" with j to withdraw electrons from i. Broadly speaking, the Markov chain describes the evolution of the system (the movement of electrons around the atoms in this case) in two different scales, the "short term" and the "long term". In the short-term scale of time (first interval of time, t_1–t_0=1) the random movement of electrons is described by ${}^{1}\Pi$, whilst long-term movements are described by the Chapman–Kolgomorov equations:

$$ p_{ij}(t_m + t_n) = \sum_k p_{ik}(t_m) \cdot p_{kj}(t_n) \tag{2} $$

In particular, it is simple to derive the relation ${}^{k}\Pi$ ($^{k}p_{ij}$)=(${}^{1}\Pi$ ($^{1}p_{ij}$),)^{k}, which determines that the matrices whose elements are the probabilities with which electrons move from atom i to atom j in time t_k ($^{k}p_{ij}$) are the kth natural power of ${}^{1}\Pi$ ($^{1}p_{ij}$). [1, 2, 3, 56, 58] Figure 2 shows an example for the calculation of short-term probabilities that will be explained later on in this section.

It does not make any difference if the Pauling scale ($^{1}\chi_j$) or any other linearly related scale ($^{\Pi}\chi_j$=a·$^{I}\chi_j$) such as Kier–Hall electronegativity [59] is selected. In fact, the present approach is invariant to the selection of the electronegativity scale:

$$ {}^{1}p_{ij}\left({}^{II}\chi\right) = \frac{{}^{II}\chi_j}{\sum_{k=1}^{\delta+1} {}^{II}\chi_k} = \frac{a \cdot {}^{I}\chi_j}{\sum_{k=1}^{\delta+1} a \cdot {}^{I}\chi_k} = \frac{a \cdot {}^{I}\chi_j}{a \cdot \left(\sum_{k=1}^{\delta+1} {}^{I}\chi_k\right)} $$

$$ = \frac{{}^{I}\chi_j}{\sum_{k=1}^{\delta+1} {}^{I}\chi_k} = {}^{1}p_{ij}\left({}^{I}\chi\right) \qquad (3) $$

where the letter a refers to a constant that relates the two scales of electronegativity. It is also noteworthy that in the present approach it is not necessary but possible to use electronegativity scales that distinguish between hybrid states of atoms in bonds. For instance, sp^3, sp^2, and sp carbon have the same Pauling electronegativity but are clearly distinguished in the present approximation (see Fig. 2). The use of other scales, not only electronegativity related, is beyond of the scope of the present study and will be considered in more detail elsewhere. In any case, the use of atom charges, charge densities, or bond orders calculated using quantum mechanics methods or semiempirical methods is time consuming, and does not offer any additional advantage. [49, 50, 51, 60]

The stochastic matrix previously described may be used to generate numerical indices of molecular structure. Here, we shall use the sum of the self-return probabilities of the natural power of this matrix (${}^{SR}\pi_k$). [49, 50, 51] In classical Markov theory, these numbers are the probabilities with which the system returns to the initial state. [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 56, 58] In the present context, they are the probabilities with which electrons return to the atoms at different times after an arbitrary initial observation time t_0.

$$ {}^{SR}\pi_k(S) = \sum_{i=1}^{g} {}^{k}p_{ii} \qquad (4) $$

The ${}^{k}p_{ii}$ are the entries in the principal diagonal of ${}^{k}\Pi$ matrixes and S is a group of atoms that compose a chemical group in the molecule. When S contains all the atoms in the molecule, ${}^{SR}\pi_k(S)$ becomes a global molecular index and we write only ${}^{SR}\pi_k$. The 0-step-self-return-electron-transition probabilities to the atom a_i (${}^{0}p_{ii}$) are the values of the principal diagonal of ${}^{0}\Pi = ({}^{1}\Pi)^0 = I_n$, where I_n is the identity matrix of order n. Therefore, ${}^{0}p_{ii}$ is, by definition, equal to 1 for any atom, and ${}^{SR}\pi_0$ is equal to the number of atoms in the molecule. This fact has a simple physical meaning: at time 0, electrons can do only one thing: obviously, to stay around their atom with probability 1. The calculation of ${}^{SR}\pi_k$ for any organic or inorganic molecule was carried out using the MARCH-INSIDE software. [61] This software has a graphical interface to make the chemist's work easier (see Fig. 3).

In Fig. 2, we exemplify the definition and calculation of the ${}^{1}\Pi$ matrix for nitrilo-acetyl fluoride. This molecule contains five atoms, thus ${}^{0}\Pi = I_5$. Therefore, by definition, ${}^{SR}\pi_{0=}\mathbf{Tr}$ (${}^{0}\Pi$)=5. The symbol **Tr** represents the mathematical operator Trace (sum of the entries in the principal diagonal of the matrix). [45, 46, 49, 50, 51] From ${}^{1}\Pi$ and ${}^{2}\Pi$ we can calculate ${}^{SR}\pi_{1=}\mathbf{Tr}$ (${}^{1}\Pi$)=2.443 and ${}^{SR}\pi_{2=}\mathbf{Tr}$ (${}^{2}\Pi$)=2.143.

In more detail, it is also shown that ${}^{1}p_{ii}$ varies in the following order: ${}^{1}p_{ii}(F)=0.615 > {}^{1}p_{ii}$ (O)=0.583 > ${}^{1}p_{ii}$ (N)=0.455 > ${}^{1}p_{ii}$(C2)=0.313 > ${}^{1}p_{ii}$(C1)=0.200. We may conclude that ${}^{1}p_{ii}$ varies in the same order as the electronegativity (χ_F=4.0 > χ_O=3.5 > χ_N=3 > χ_C=2.5). It is obvious that electrons will have a higher Markovian probability of returning to the sp carbon (0.313) than to the sp^2 carbon (0.200) despite using the same electronegativity. This fact is in line with quantum mechanical results, the electronic density around linear (sp) carbon atoms is greater than in sp^2 carbon atoms, and may have important implications in QSAR. [62] We may argue this good differentiation of the atoms with different hybridization if we consider the topological character of ${}^{1}\Pi$. As shown in Fig. 2, both ${}^{1}p_{ii}$(C2) and ${}^{1}p_{ii}$(C1) have identical numerators (χ_C), but different denominators. This occurs due to the different "connectivity" of the two atoms, i.e., C1 is connected to O, F, and C2 while the C2 atom is bonded to C1 and nitrogen.

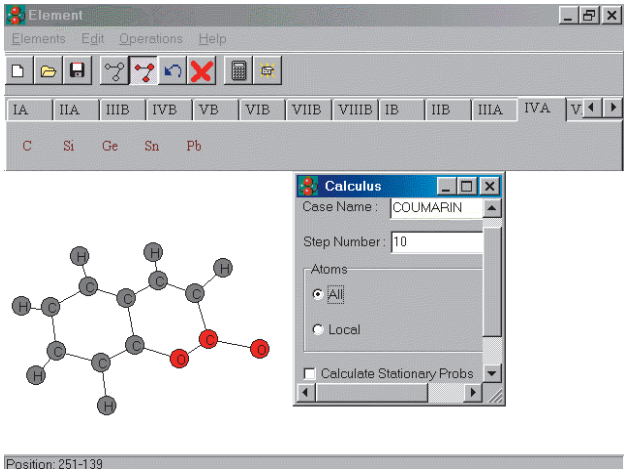

Fig. 3 Representation of coumarins' basic core in MARCH-INSIDE interface

We may therefore assert that the molecular indices (${}^{SR}\pi_0$) calculated by MARCH-INSIDE codify both electronic and topological information about molecular structure. In future papers, we will discuss this issue in more detail.

The use of the symbol **Tr** clearly shows that the present molecular descriptors are formally the spectral moments of ${}^{k}\Pi$. Spectral moments of other structural matrices have also been studied in the chemical literature over a long period, in diverse chemical contexts. [63, 64, 65, 66, 67, 68, 69, 70, 71]

Statistical analysis

Continuing from the previous section, we can try to develop a simple linear QSAR using MARCH-INSIDE methodology using this general formula:

$$ \mathrm{ACA} = b + b_0^{SR}\Pi_0 + b_1^{SR}\Pi_1 + b_2^{SR}\Pi_2 + \ldots + b_k^{SR}\Pi_k \qquad (5) $$

Here the structure is represented by the molecular indices ${}^{SR}\pi_k$ and the activity (anticancer activity in this case) by the variable ACA (acronym of anti-cancer activity). This is a dummy variable, ACA=1 for anticancer compounds and ACA=−1 for the non-active compounds. In Eq. (5) b_k are the coefficients of the classification function, determined by least squares, as implemented in the Linear Discriminant Analysis (LDA) module of STATISTICA 6.0. [72] Forward stepwise was established as the strategy for variable selection. [72, 73, 74, 75, 76]

To develop the QSAR for anticancer/non-anticancer compound discrimination, we use the first 11 ${}^{SR}\pi_k$ as molecular descriptors. The quality of the model was determined by examining Wilks' λ statistic, Mahalanobis distance, the percentage of good classification, and the proportion between the cases and variables in the equation. Calculating the percentages of good classification in the external prediction series allowed the model to be validated. Compounds in the external prediction series were never used to develop the classification function.

Here we considered general data for 961 organic chemicals that contain almost all of the anticancer chemicals reported by Negwer in its large database. [77] All the cases were processed using k-Means Cluster Analysis (k-MCA) in order to design predicting and training data series. Firstly, we carried out a k-MCA1 with the active compounds and later another, k-MCA2, using the inactive compounds. Anticancer and non-anticancer training series were selected at random (214 active and 467 inactive compounds). The remaining sub-series was used as an external prediction series, containing 84 anticancer and 196 non-anticancer chemicals. Figure 4 graphically illustrates this procedure.

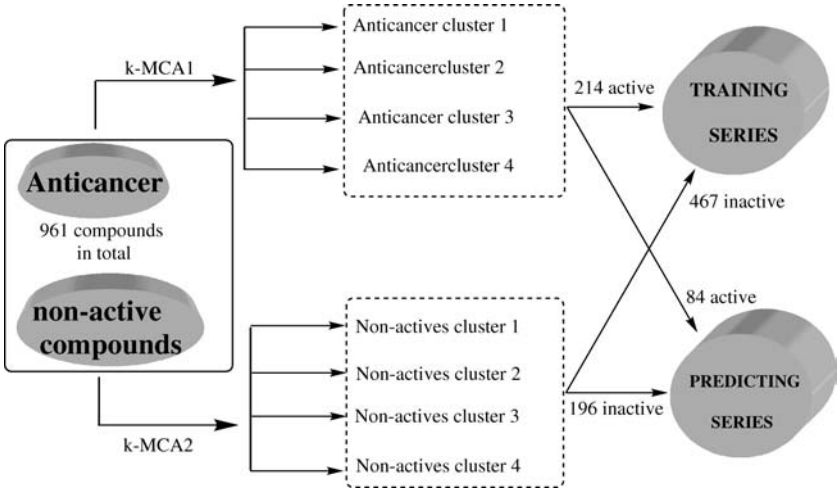

Fig. 4 General algorithm used to design training and predicting series

The *k*-MCA was carried out with the same software as LDA, but using the *k*-MCA module. For acceptable statistical quality of data partition in clusters, we took into account the number of members in each cluster and the standard deviation of the variables in the cluster (as low as possible). We also inspected the *between SS* and *within SS* (Standard deviation between and within clusters), the respective Fisher ratio and their *p*-level of significance considered to be lower than 0.05. [78, 79]

Iso-contribution zone analysis (IZA) and MARCH-INSIDE

In order to calculate the total atom contribution to anticancer activity in the current approach, we make use of the decomposition of total molecular descriptors into local descriptors. More specifically, we decompose the total molecular descriptors into atomic descriptors of the atom in the molecule. For example, the molecular descriptors of chloroform may be decomposed as follows: ${}^{SR}\pi_k(HCCl_3)={}^{SR}\pi_k(H)+{}^{SR}\pi_k(C)+3{}^{SR}\pi_k(Cl)$. Afterwards, the values of the atomic descriptor for each atom are substituted in the QSAR equation, obtaining the contribution of the atom to anticancer activity. Estrada and González have recently explained this procedure in detail for bond spectral moments. [46]

A step forward in this regard was offered by some of the authors of the present paper, by regrouping all positive (negative) contributions in order to obtain a picture that maps the molecular regions with positive or negative contribution to the property. The method, called Iso-Contribution Zone Analysis (IZA), is general for any molecular descriptor, defined a priori as a sum of local descriptors, at least for linear QSARs. [52] The main importance of IZA is that it offers a clear and direct interpretation of results in structural terms. Here we adapt an IZA approach to MARCH-INSIDE and LDA methodology. The present study is aimed at the selection of novel drug candidates for synthesis. Then, we select the different structural synthetic blocks of the molecules as molecular regions for the IZA. As LDA predicts the probability of action, we preferred to standardize all of the contribution in order to express them as the percentage of activity that each group accounts for.

Biological activity

Cell cultures

Human Myeloid Leukaemic Cells (HL-60) were grown in RPMI 1640 (Sigma Chemical Co.) supplemented with 15% heat-inactivated fetal calf serum (Seromed). Human Cervix Adenocarcinoma Cells (HeLa) were grown in a nutrient mixture F-12 [HAM] (Sigma Chemical Co.) supplemented with 10% heat-inactivated fetal calf serum (Seromed). Up to 100 U ml^{-1} of penicillin, 100 μg ml^{-1} amphotericin B (Sigma Chemical Co.) were added to the media. The cells were cultured at 37 °C in a moist atmosphere of 5% carbon dioxide in air.

Inhibition growth assay

HL-60 cells (3×10^4) were seeded into each well of a 24-well cell culture plate. After incubation for 24 h, various concentrations of the test agents were added to the complete medium and incubated for a further 72 h. A similar treatment was used for HeLa cells (see for instance [80]). A trypan blue assay was performed to determine cell viability. Cytotoxicity data were expressed as IC_{50} values, i.e. the concentration of the test agent inducing 50% reduction in cell numbers compared with control cultures. UV sample irradiation was performed using a Philips HPW 125 (365 nm). The intensity of radiation (14.075 mW cm^{-2}) was determined with a Cole–Palmer radiometer (model 97503-00). All chemicals (analytical degree) were purchased by the Department of Organic Chemistry in the Faculty of Pharmacy at the University of Santiago de Compostela, Spain.

Results

The *k*-MCA was used in the design of training and predicting series. It allows us to design both training and predicting series that are representative of the entire "experimental universe". We first carried out a *k*-MCA with 298 anticancer compounds and afterwards with 663 non-anticancer compounds. The first analysis yielded clusters of active compounds and the second the same number of clusters of non-active compounds. The variables ${}^{SR}\pi_0$ to ${}^{SR}\pi_3$ were used, with all variables showing *p*-levels of <0.05 for the Fisher test. The results are shown in Table 1.

Once the random and representative selection of training series is carried out, it is possible to fit the discriminant function. The QSAR-LDA model selection was subjected to the principle of parsimony. We then chose a function with high statistical significance, but with as few parameters (a_k) as possible: [81]

$$\mathrm{ACA} = 3.032^{\mathrm{SR}}\pi_0 - 49.519^{\mathrm{SR}}\pi_1 + 126.634^{\mathrm{SR}}\pi_2 - 165.795^{\mathrm{SR}}\pi_4 + 215.591^{\mathrm{SR}}\pi_8 - 128.236^{\mathrm{SR}}\pi_{10} - 6.579$$

$$N = 681 \lambda = 0.443 F = 141,31\ D^2 = 5.841 p < 0.00 \quad (6)$$

Here, λ is Wilks' statistic, which for overall discrimination takes values in the range from 0 (perfect discrimination) to 1 (no discrimination). Comparison between Mahalanobis distance (D) and Fisher ratio (F) allows us to check the hypothesis of separation of groups with a probability of error (p-level) of $p < 0.05$.

This model correctly classified 90.5% of the compounds in the training series, i.e., 65 misclassifications in 681 cases, while in the predicting series there were 39 errors in 280 cases, i.e. 86.1% of good classification. More specifically, the model correctly classified 90.2% of anticancer compounds in training series and 84.5% of these compounds in predicting series. The classification results and the names of each anticancer compound used in both training and predicting series are shown in Tables 2 and 3.

In these tables (2 and 3) and the others, $\Delta P\% = [P(\text{actv}) - P(\text{non-actv})] \times 100$, where P (actv) is the a posteriori probability with which the model classifies a compound as active. Conversely, P (non-actv) is the a posteriori probability with which the model classifies a compound as non-active. This value ($\Delta P\%$) takes positive values when P (actv)>P (non-actv) and negative otherwise. Therefore, when $\Delta P\%$ is positive (negative) the compound was classified as anticancer (non-anticancer). When $\Delta P\%$ was in the range $-5 < \Delta P\% < 5$ the compound was considered as unclassified. A P(actv)x100 value higher than 50 is considered as a threshold limit to classify a compound as highly active, although we prefer to use a stronger criterion, $\Delta P\% > 50\%$. [49, 50]

Elsewhere, the model correctly classified 90.6% of non-anticancer compounds in training series and 86.7% of these compounds in predicting series. The classification results and the names of each non-anticancer compound used in both training and predicting series are shown in Tables 4 and 5.

Our research groups have been involved in the in vitro search for anticancer compounds. [80] Special emphasis has been given to the search for *n*-methoxypsoralen (*n*-MOP) derivatives. [82, 83, 84] In order to test the potential of MARCH-INSIDE and LDA for detecting novel anticancer leads, we predicted the biological activity of all the chemicals contained in a combinatorial library of coumarin derivatives. The library contains drugs-like chemicals with the most common substituents in medicinal chemistry, [85] attached at all positions of coumarins' core, as well as condensed cyclic derivatives. We then selected a group of four chemicals (see Fig. 5), among those with higher probability of anticancer action, to be tested in an in vitro antiproliferative assay (see Table 6).

Finally, we applied IZA in order to carry out an interpretation of the classification function in structural terms. The IZA picture for one anticancer compound is depicted in Fig. 6. As was explained in the Materials and methods section, zones shown in black (shown in white) are those that have a negative (positive) contribution to anticancer activity.

Table 1 Results of the K-Means cluster analysis

Anticancer compounds								
Cn/Nc[a]	1/108	2/85	3/62	4/43	Global cluster statistical analysis			
Variables	Standard deviation of clusters				SSb[b]	SSw[c]	F[d]	P[e]
$^{SR}\pi_0$	4.43	3.47	4.18	6.97	54112.9	6213.9	853.4	0.00
$^{SR}\pi_1$	1.45	1.14	1.32	2.34	5400.3	669.1	790.9	0.00
$^{SR}\pi_2$	1.09	0.82	0.93	1.80	2685.2	371.1	709.0	0.00
$^{SR}\pi_3$	0.98	0.72	0.80	1.60	1988.0	293.2	664.5	0.00
Non-anticancer compounds								
Cn/Nc[a]	1/262	2/173	3/157	4/71	Global cluster statistical analysis			
Variables	Standard deviation of clusters				SSb[b]	SSw[c]	*F*[d]	*P*[e]
$^{SR}\pi_0$	5.11	2.49	2.89	3.41	70009.6	9992.6	1539.0	0.00
$^{SR}\pi_1$	1.62	0.76	0.90	1.05	6428.3	991.6	1424.0	0.00
$^{SR}\pi_2$	1.13	0.53	0.64	0.76	2993.8	487.9	1347.9	0.00
$^{SR}\pi_3$	0.96	0.45	0.56	0.66	2153.1	356.6	1326.2	0.00

[a] Cn/Nc=cluster number/number of cases in this cluster
[b] SSb=SS between
[c] SSw=SS within
[d] *F*=Fisher ratio
[e] *P*=signification level

Table 2 Results of discriminant analysis for anticancer compounds in the training series

100≥Δ*P*%[a]>99				
EMDAI	Azotomycin	Diaziquone	Meturedepa	AB 100
Nannosulfan	Spergualin	Crotepoxide	Mitoxantrone	Benzodepa
Etoglucid	Dehespamine	Sibiromycin	Estreptozocil	Rufocromomycin
Dipin	Magnnityl Dimesilate	Diazan	Neplanocin c	Hexaphosphamid
Mesyldegranol	Mannomustine	Aphoxide	Asalei	Psicofuranine
Inproquone	triaziquone	AB-182	Hexestrol $(PO_4)_2$	Asaline
A-139	Ketotrexate	Azaserine	Mitobronitol	Rabdophillin G
Fotetramine	Pidorubucin	Teralphezin	Tretamine	Stibostat
Alazopeptin	Rutin-N-Mustard	Pteropterin	Chlorozotocin	Fludarabine
Cervicarcin	Tetracicline	Carboquone	Bactobolin	Methopterine
Amygdalin	Hesperidin	Idarubucin	Toromycin	Astiron
Medorubicin	Pactamycin	Solafalmitin	Menogaril	Leatrile
Dauronobicin	Withaferin A	AT-16	A-Ninopterin	
99≥Δ*P*%[a]>90				
Lomenin-2	Eupochlorin acetate	Prosfidium chloride	OPSPA	Sangiuamicin
Dimetfolamide	Benzotef	Fosfemid	Irisquinone A	Benaxibine
Duazomycin	Aminopterin	Chlorbutifenicillin	ODEPA	Disulfbumide
Fluorbensotef	Diiodbenzoteph	Amino Anfol	Porfiromycin	Estramustine
Metamelfalan	Aminotreofol	Phansazin	Estramustine PO4	Lysopsin a
Mitopodozide	Hexadepa	Uramycin	Amebisan	Defosfamide
Ambunol	Dinaphtimine	Pirabofurin	Cleisthathin	Triciribine
Bremfol	CAM	Chlorasquin	Fludarabine	Cytaracid
RPCNU	Ditiomustine	Fluorasquin	Leucodelphinidin	Phenamet
M-83	Mitomycin	Thioguanosine	Araside	
Octostanolon	Methasquin	Nitrocafan	Ara-T	
Diethylstilbestrol (SO4)2		**Acetoxycycloheximidine**		
90≥Δ*P*%[a]>60				
Fentirin	Citarabine	Asperlin	Hisfen	Trimetrexate
Thioinosine	Benzolide	2-AA	Lofenal	CB-10252
Formicyn	Fluoromezin	Osayin	Ocaphane	Chlorambucil
Alalon	IDA	Merophan	Damuar	Fenastezin
Lysopsin b	Bututricin	Piposulfan	Fluorafur	GEA-29
Azazipidine	Aldophosphamide	Drostanolone	Asazol	Forfenimex
Isopropylcad	Glutacyt	Thiazofurine	Promicil	V-100
CB1837	Sparzomycin	Dichloroallillawsone	Dimezol	Blueidon
Butastezine	Hidroxycicloheximide	Butoctamide	Athoxen	Nifuron
60≥Δ*P*%[a]>5				
Sparzomycin	Busulfan	Doxifluridine	Butodicin	IOB-177
Spirazidin	Ac. Mycophenolicum	Juncusol	Cyanocyline A	Genirin
BA1	Piritrexim	Alanosine	Flurocitabine	Neptamustine
Bufloracil	Chlorphenacil	Angustibalin	Piperazinedione	Lumostine
Nimustine	Cafencil	Spiromustine	Semustine	
Ripazepam	Trestolone acetate			
Misclassified compounds (−5≥Δ*P*%[a])				
NSC-83265	Tylophorine	Leucenol	QFI	Nitracine
Burseran	Oxymatrine	Homocoralyne	NSC-95466	Enterolactone
Testolactone	Leukogen	Vinervine	Butocin	Zimet 54/79
Benzatine	Aceglatone	Bimolane	Spirogermanium	Bisantrene+A239
Laveldamycine	Peucedanin			

[a] See explanation in the text

Discussion

Due to differences in the composition of experimental data and the method used in carrying out the QSAR, it is not feasible to carry out a comparison between the models reported in the literature for the selection of anticancer compounds. In fact, almost all-anticancer activity QSARs are based on homologous series (specific families) of organic compounds. [35] In any case, for screening purpose it is obviously more useful to use comparable data obtained by general and not class-specific models. In addition, the chemical classes of the training compounds limit the applicability domain of the above-mentioned models. [86] We then selected a previous model reported by our group using the TOPS-MODE approach. [87] The selection is based on the use of LDA as a method for deriving the QSAR, the important diversity of chemical structural patterns contained in the data, and the use of the same source for collecting the data.

The percentage of false actives obtained in the training series was higher than that reported for the TOPS-MODE approach. The previous study reported a 5.0% of false

Table 3 Results of discriminant analysis for non-anticancer compounds in the training series

$-100 \leq \Delta P\%^{a} < -95$				
Phenindamine	Naphasoline	Clotiapine	Phensuximide	Fluperlacine
Kepone	Brosuximide	Clozapine	Mianserin	Dichobenil
Mirex	Dichlone	Acetanilida	Phenmetrazine	C-56
Aldrin	Isobenzan	Tetrahydrozoline	Chlordane	Lofemizole
Tetrachlorothiophene	Amoban	Parathiazine	Pyrazon	Antipyrine
Selectan	Morestan	Pyrathiazide	Azanator	Chorcyclizine
Pentachlorophenol	Phenindione	Amitrole	Cycliramin	Naftoclizinum
p-Dichlorobenzene	Fenharmane	KB1043	Diphenylhidantoin	Dyrene
Phenothiazine	Chloranil	Phenylhidrazine	Methdilazine	CBZ
Dehydroclothepine	Strycnine	S. 131	Ammonium sulfamate	
$-95 \leq \Delta P\%^{a} < -90$				
Perathiapten	CyElizine	Arecoline	Metasuximide	Anphetamine
Phensoximide	Foeirtoline	Ovex	Dibenamine	Pargyline
Monuron	Fenuron	Zotepine	Paracetamol	Cinromide
Zoxazolamine	Tetradifon	Mycocid	Linuron	Methylene blue
Acetophenetidine	Meclizine	Picartamide	Caffeine	AB 41
Metazide	Diuron	Benzoctamine	Metacetamol	Naranol
Chrordiazepoxide.	Ethyllisergamide	Metipirox	Desipramine	Acrolein
Iniazid	Ethosuximide	Heptauerine	Prothixene	Metaxalone
Methan sodium	Pyroxamine	Nicotafuryl	Phenacemide	
Antu	LD2855	Folpet	Tiquinamide	
Desmethylprothiaden		Naphtalene Acetic Ac.		
$-90 \leq \Delta P\%^{a} < -85$				
Methoilazine	Amezepine	Tifemoxone	Emorfazone	MCP
Clorprothixene	Chlorphenacemide	Sulfanilamide	Bromoxynil	Pentylenetetrazol
Tolexantone	Maneb	Pentanal	Midamaline	Stenofril
TDE	Prochlorperazine	4-aminophenasone	Mephenoxalone	Serotonin
Mexamin	Imidan	Ethenzamide	pyrolan	Paracrofamol
Tolindate	Methanphetamine	Bromamide	Prooxen	Equilenin
SU-7692	Brusine	Nikethamide	Aezulanum	Isocarboxazid
Genite	Glycopyramide	Allisan	L11204	Glutethimide
Methapyrilene	Amethobenzepine	Dichloran	Tripelennamine	Methetoin
Fantridone	IB 503	Clomacran PO4	Trimethadione	
$-85 \leq \Delta P\%^{a} < -80$				
Sulfapyridine	Demexiptiline	B 777-81	Banol	Thiogin
Salicylamide	Clodazon	Metaclazepan	Butylparaben	Isotiquimide
Hydrastinine	2,4,5-T	Sulfadiazine	Orphenadrine citrate	SOG-18
Doxofylline	Thebaine	Fluoroacetate	Phenylbutazone	Sulfacetamide
PRL 8–53	DDT	Dibrosalicyl amide	Fosazepam	Chlomethizole
Difencloxazine	Oxyphemedazol	Cycloterenol	Tolpropamine	Arcylate
Etofuradine	Chrormezanone	Heliofilm	Mephenoxalone	Methapyrilene
2.4-D.	Thiadrine	Primidone	Dalapon	Diethazine
Methamilane	Ioxynil	Captan	Paramethadione	Chlordinezin
Benzamsulfonium	Mephentermine	Pirprophen	Diclonixin	
$-80 \leq \Delta P\%^{a} < -70$				
Diethazine	Lindane	CIPC	HCA	Haloperidol
Sulfinpyrazone	Dexon	Epirizole	Acepromacine	Herbisan
Isothipendyl.	Dacthal	Pibenzepine	Promazine	Pipradrol
Cycrimine.	BW775C	Barbital	Carbophenithion	Beztiacide
Acetergamine	Neburon	Silvex	Dimethoate	Dimetilan
Thimerosal	Chlorothen citrate	Hydrochlorotyazide	Mesocarb	Phenyramidol
Equilin	Metofurone	Picloram	Meflophenhidramine	Trenbolone
Ziram	Ronnel	Solan	Almoxatone mesilate	Ectylurea
Trichloroethanol	Hydrocodone	Dimezin	Acylmidrazone	Vorhexobarbital
Alimemazine	Aminopyrine	Cyclopentamine	Diethylglycyl phenothiazine	
$-70 \leq \Delta P\%^{a} < -60$				
Cyclobarbital	Simazine	LY 125180	Mecloralurea	Difenbutamine
Pheneturide	Bifemelane	Neurodin	Iproclozide	Apazicin(A)
Crotamiton	Sulphaethylpirazole	Metharbital	T28	Thebaine
S1688	Duraseries	Bou-14607	Ciglitazone	2.4-DB
Orphenadrine	Atrolactamide	Ethylmorphine.	Ac. Nicosalicylicum	Matacil
Ro 11–4337	Pyrilamine	Giareg	Bromaspirin	Pectol

Table 3 (continued)

Thonzylamine	Ceresanim	Salycilic acid	Zingeron	Amendol
Phorate	Nialamide	Methyl demeton (B)	Biperiden	Carbromal
Aclu	Difolatan	Nitrofurazone	Ferban	Promethazine
Meperidine	Dichrorphenamine	Perthane	CP15525	
NH_2Phenylamidophenazone				
$-60 \leq \Delta P\%^a < -40$				
Dilan	Phenyl-propanolamide	Uracil mustard	Nitrofurylen	Isolan
Neostigmine Br	Dimazenum	Dicofol	Phenacaine	Lidocaine.
Podilfen	Sulfadicramide	Sesone	Disulfoton	Baygon
Estrone	Heptabarbital	Ametryne	Randox	Ephedrine
Physostigmine	Reseran 13	EF-525	Moxastine	Dursban
Carbachol	Piperalin	Ethoheptazine	Metopon	F-28
Mefenamic Acid	Bromocyl	Mescaline	orthouanizide	Vinbarbital
Votracon	Xenyhexenic Acid	Dimethisoquin	Anitrazafen	Imipramine.
Levallorpran	Chlorphenoxamine	Embramine	Zectran	Doxilamine
EpTAM	Morphine	Chlorobenzilate	Alclofenac	
Mepivacaine	MGK repellent 11	Oxyphenbutazone	Brindoxime	
Succinic acid 2,2 dimethylhydrazide		Pyrilamine maleate		
$-40 \leq \Delta P\%^a < -10$				
Zytron	Warfarin	Nitroxazepine	Allobarbital	Chlorphenesin
Tolazamide	TR 35	Hexacaine	EPN	Amobarbital
Dyclonine	Dimethoxanate	Mepensolate	Cocaine	Dibatod
Pipazethate	Prilocaine	Heroin	Tricyclamol	Aminoprofen
Bisacoryl	Temik	Trifluoperazine	Arsthinol	Atratone
Coumachlor	Vernam	Migyl	Oxabrexine	Diampromide
Thionazin	Phenoxybenzamine	Homo-Pas	Dasanit	Methoxychlor
Papaverine	Pebulate	Methyclothiazide	Hydroxizine	Piperidolate
Dyclonine	Buclosamide	Metaraminol	Butaverine	Bupiracaine.
Diphenizin	Metochalcone	Meparfynol	Fludorex	Chlorthion
Oxomezazine	Cycloguanil	Norclostebol	Synafinamide	Chrorphenesin
$-10 \leq \Delta P\%^a \leq -5$				
Propiomazine	Disulfiram	Piperocaine	Hexylcaine	Homococaine
Thiopental	Phenadoxone	Bufivacaine	Hexobarbital	
Misclassified compounds ($\Delta P\%^a < 5$)				
Aminoteropterin	Isoproterenol	Chloranbucil	Amedin	Nitrendipine
Carisoprodol	Methohexital	Acetylcarsonobenzol	Polytiazide	Colchicine
Butacaine SO4	TABAC	Alifedrine	Flumethyazide	Picrotoxin
TU 399	Ciamexon	Dihydroergotamine	Carazolol	Pentapiperide
Dibucaine	Probenecid	DEF	Morphine	Glicerothiazol
Isopentaquine	Mariptiline	Prometone	Bishidroxycouma-rina	Procaine
Meprobamate	Caramiphen	Diphenhidramine	Oxymetazoline	Rexamid
Atropine	Naepaine	Aminohexan	Fluoroquine	Prometryne
Talbulal	Choumaphos	Propazine	Gobab+A286	
Non-classified compounds ($-5 < \Delta P\%^a < 5$)				
Lilly 51641	Clofencilan	Carbutamide	Demeton-O	ASA
Chroroquine PO4	Cefaloram			

[a] See explanation in the text

actives, while our present model misclassifies 9.5% of the compounds in the training series. Both values are generally very good, if we consider the broad spectrum of chemical structures, although we should remember that the model reported here uses a data series three times larger than that used in the former, i.e., 681/224 compounds. Another important factor is the results of the classification in predicting series with regard to training series. It is reasonable to expect some decrease in overall predictability of predicting series with respect to training series for a simple reason; the model is developed to fit the points in training series, and therefore data points in predicting series are never used to develop it. Our previous model [87] has shown higher classification percentages in predicting than in training series. It could be determined by a not exactly random selection of both series. In the present work, the use of *k*-MCA to design training and predicting series effectively overcomes this problem. In any case, the results in predicting series fully validate both models, for practical use, from a statistical point of view. [88]

As previously indicated, our research groups have mainly worked on trial-error searching for anticancer compounds. [80, 82, 83, 84, 87, 89, 90, 91] At the same

Table 4 Results of discriminant analysis for anticancer compounds in the predicting series

$100 \geq \Delta P\%^{a} > 95$				
Teroxirone	Lonin 4	Ritrosulfan	Pumitepa	Mitolactol
Thiodirin	Ganu	Denopterin	A-Denopterin	Thiohexadepa
Diopterin	Epipropidine	Etofoside	Triciribine PO_4	Loglutam-2
Iremycin	Euparotin acetate	Ederpin	Bufumustine	Asamet
GYKI 13324	Elderfiel pirimidine	Azatepa	Calcii Mefolinas	Methotrexate
Holacanthone	Cytochalasin B	Asdofan	Don	1954CD
Esorubicin	Ribofrine	Ac. Sparfosicum	Benzodet	Trichodermin
$95 \geq \Delta P\%^{a} 80$				
Quinaspar	Gliocadic Acid	Fenafan	Neplanocin	Inprosulfam
Marcophan	Phenaline	Nicosin	Bendamustine	Magestrol
Indicine *N*-oxide	Aethimidinun	3-Deazaguanosina	C61	Calusterone
Enterodiol	Aminochlorambucil	Aminoalanfol	8-MOP	Dopastin
$80 \geq \Delta P\%^{a} 10$				
Triazinate	G-azauridine	Trophosphamide	Demecolcine	Auxitabine
Sulfofamide	Alkyron	Elmustine	Aziprin	IMET-3995
Alanine Bustard	Citostal	Macaine	Phenester	
Non-classified ($5 \geq \Delta P\%^{a} \geq -5$)				
IMET-3106	Metfol-B	Pseudourea		
Misclassified ($-5\ 5 > \Delta P\%^{a}$)				
Mitoguazone	Coralyne chloride	Azathioprine	*m*-Embitol	CRC-7001
Enpromate	*o*-Embitol	*p*-Embitol	Albonoursin	Hainanolide
CGP-15720	BRL-51308			

[a] See explanation in the text

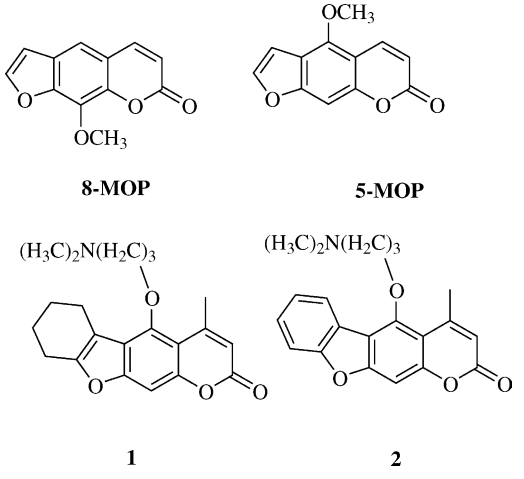

Fig. 5 Chemical structures of the assayed chemicals

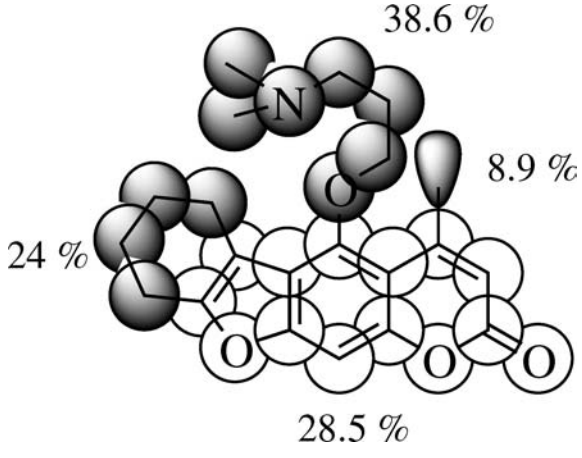

Fig. 6 IZA of compound **1**

time, virtual screening (based on QSAR techniques) has emerged as an interesting alternative to high-throughput screening. [19, 87, 92, 93] Here we perform "in silico" mining into a combinatorial library of coumarins looking for novel anticancer compounds by using the discriminant function obtained through the MARCH-INSIDE and LDA methodology. The results shown in Table 6 exemplify how the present approach could be used for the selection of possible anticancer drug candidates. All chemicals in this table were predicted with $\Delta P\% > 90$. This table shows the results of the "in vitro" tests for these coumarins and two control drugs. In general, psoralens (linear furocoumarins) have been mainly known as UV-light-activated antiproliferative compounds. [94] In particular, both the UV-induced or in-darkness activity of 8-MOP has been the subject of increased research interest. [95] As shown in Table 6, both psoralens (**1,2**) presented similar to higher activity than 8-MOP and 5-MOP in the presence of UV light. It is noteworthy that both compounds **1** and **2** also have antiproliferative activity in the darkness. In any case, chemical **1** had the highest activity. It is interesting to note that both chemicals present the $-O(CH_2)_3N(CH_3)_2$ substituent. Our group has recently discussed the favorable effect over biological activity of this group in other families of compounds. [80] Furthermore, other authors have reported the use of this structural pattern as a linking functions ($-(CH_2)_2N(CH_2)_2-$) to increase the biological activity of anticancer compounds. [96]

Table 5 Results of discriminant analysis for non-anticancer compounds in the predicting series

$-100 \leq \Delta P\%^a < -90$				
Maleic hidrazide	Strinoline	Mebicar	Oxazepam	Methaqualone
Diphenyl	Azamianserin	Dimemorphan PO4	Oxasepam	Carbaryl
Deximafen	Loxapine	Praxadine	Ciclopramine	AcKet
Heptachlor	Pyrantel	Aethosucxinid	Dicryl	Metane arsonate
Dimethylsulfoxide	Chlorbenside	Metacetanilidum	Sirmate	Fenac
Nicotine	Sweep	Ethotoin	Metasuximide	Bemegride
Basedol	Histamine			
$-90 \leq \Delta P\%^a < -80$				
Mephenytoin	Thiram	Thenyldiamine	Hydrocodone	SNF 70948
Indopan	Mezepine	Vinconate	Phenobarbital	Promethazine
A 29 Lundbeck	VOFP-12392	Sulfathiazole	Methyl salicylate	Dimefox
Methyl Trithion	Molinate	Amiben	Azaprocin	Apomorphine
Zineb	Deoxyestrone	Furazonal	Diclofenac	
$-80 \leq \Delta P\%^a < -70$				
Eusolex 8020	Tandamine	Zenbromal	Picloram	Dicamba
Chorothiazide	Trimeprazine	Bromacil	Dibenzepin	Oxadimedine
AL-1965	Barban	Proquazone	Mesurol	Vegadex
IPC	Diethil carbazine	Paraidehyde	BT 132 Merck	Brosotamide
Salinazid	Beloxamide	Tripelennamine	LSD	Clormecaine
$-70 \leq \Delta P\%^a < -50$				
Tranilcypromine	Dieldrin	Metifenazone	GC-6,506	Codeine
Diamide	Methanarsonate	Asa	Dalapon salt	Isometheptene
Methopromazine	Methanarsonate Ac.	Rodocaine	Nabam	Brofezil
Nonaferone	*m*-Methoxythioacetazona	Phenkapton	Nalidixic acid	Ro-Neet
Guayacol	*p*-Methyldiphenhydramine	Alanap	Tizolemide	Heptabarbital
Pinafide	Phenthimentonium	Endrin	Sulfaguanidine	Oxycodone
Sytramate	Doxapram	Nemacide	KF 1492	Phenilephrine
$-50 \leq \Delta P\%^a < -30$				
Benzazetin	Sulfamethizole	Bisacoryl	Sulfatroxazole	Medrylamine
Methyl demeton (A)	Racefemine	Sulfadimethoxine	Warfarin	Glibornuride
Codein	Chloral hidrate	Closiramine	Sulfisoxazole	Sulfamoxazole
B-Aminosalicylic	Brodimofrin	Atrazine	Anot	Dihydrocodeine
Metopon	Chrormerodrin	Sulfaforazole	Sulfatriazine	Fostedil
$-30 \leq \Delta P\%^a < -5$				
Pronilide	Captodiamine	DDVP	Vraton	Cibenzoline
Clocanfamide	Rotenone	Ethamivan	Dienestrol	Cyclomethycaine
Valnocyamide	Fenetylline	Felodipine	Tetracaine	Thiolactomycin
Methyldesorphine	Methyl parathion	Methonalide	Aprobarbital	MCN-2840
Procaine	Triflupromazine	Bupiracaine	Aspamin A	Quinidine sulfate
Non-classified ($-5 \leq \Delta P\%^a < 5$)				
Ethinamate	Trichlormethiazide	Naled	Atolide	
Misclassified ($\Delta P\%^a < 5$)				
Thiphenamil	Succinylsulfathiazole	Thiopental	Carbetapentane	K4423
Eunesine	Phtalylsulfathiazole	Pentaquine	Fluphenazine	2,4 DEP
Diethylstilbestrol	Meparfynol carbamate	Methocarbamol	Tiodazosin	Methadone
Phenaglycodol	Hydroxichloroquine	Pentobarbital	Pipenzolate Br	Emetine
Pramoxine	Trimethobenzamide	Di-Allate	Cefaloglycin	Valethamate Br
Diclofurime mesilate				

[a] See explanation in the text

The IZA (Fig. 6) of **1** coincides with the facts detailed above. The psoralenic core accounts for the higher proportion of activity in the molecule (28.5% in **1**). On the other hand, the insertion of the $-(CH_2)_3N(CH_3)_2-$ group in the molecule increase the activity too (38.6%). It was previously discussed that the present group may increase either drug solubility or DNA-linking properties with the subsequent increase in activity. [80, 96] Whatever the case, the psoralen structural feature has an important positive contribution to activity in both cases. Based on the premises of the present model, this indicates that the movement of electrons in the psoralen system is largely determinant for anticancer activity. This interpretation is in agreement with numerous experimental results that have made it possible to postulate a covalent DNA–psoralens interaction, which determines the photobiological activity of psoralens. [97, 98, 99]

Table 6 Results of the biological assay

Compound[a]	IC_{50} (μM) [b]			
	Hela		HL-60	
	Darkness	UV	Darkness	UV
8-MOP	>20	10.0±3.0	>20	5.4±0.7
5-MOP	>20	16.3±0.8	>20	3.4±0.4
1	>20	1.1±0.3	5.3±2.1	0.5±0.3
2	>20	3.7±0.2	15.8±3.6	3.4±0.4

[a] See Fig. 5 for the chemical structure of these compounds
[b] See Materials and methods section

In conclusion, the development of more timely and flexible theoretical methods will lead to a new age of virtual drug discovery. [100] In this context, we may assert that the MARCH-INSIDE methodology offers a novel option for developing anticancer discovery directed QSAR in a fast and efficient way. The definitions given here could be generalized to other biological activities in order to extend the applications of MARCH-INSIDE. It is important to emphasize that this approach, together with several others, could be interpreted in structural terms using IZA.

Acknowledgements We would like to offer our sincere thanks to the two unknown referees and the editor for their critical opinions about the manuscript, which have significantly contributed to improving its presentation and quality. González DH would like to express his thanks to Dr. Jose Luis Garcia and the Cuban Ministry of Higher Education for partial financial support and help. The same author acknowledges Dr. Kier L. B. (USA) for his kind revision of other work related to our Markovian model, and who suggested several useful ideas to us. We are also indebted to Dr. Estrada (England) for former tutorship (1994–2000) and training in computational chemistry. We are also grateful for a series of lectures given in Cuba by Dr. Gutman I., which introduced us to the study of the theory of information, and to some extent random process in chemistry. Last but not least, we would like to thank Prof. Nicolais Guevara (Mexico), Prof. Israel Queiroz (Cuba) and Dr. Jorge Galvez (Valencia, Spain) for their useful help, and the Xunta the Galicia for providing us with a grant (PR405A2001/65-0).

References

1. Markov AA (1906) Bull Soc Phys Math Kasan 15:155–156
2. Bharucha-Reid AT (1960) Elements of theory of markov process on the application. McGraw-Hill Series in Probability and Statistics. McGraw-Hill, New York, pp 167–434
3. Freund JA, Poschel T (eds) (2000) Stochastic processes in physics, chemistry, and biology. In: Lecture Notes in Physics. Springer, Berlin Heidelberg New York
4. Graepel T, Obermayer K (1998) Neural Comput 11:39–55
5. Geng J, Xu D, Gong J, Li W (1998) Int J Epidemiol 27:320–322
6. Yakovlev A, Boucher K, DiSario J (1999) Math Biosci 1:45–60
7. Vorodovsky M, Koonin EV, Rudd KE (1994) Trends Biochem Sci 19:309–313
8. Vorodovsky M, MacIninch JD, Koonin EV, Rudd KE, Médigue C, Danchin A (1995) Nucleic Acid Res 23:3554–3562
9. Chou KC (1997) Biopolymers 42:837–853
10. Yuan Z (1999) FEBS Lett 451:23–26
11. Hua S, Sun Z (2001) Bioinformatics 17:721–728
12. Hubbard TJ, Park J (1995) Proteins Struct Funct Genet 23:398–402
13. Krogh A, Brown M, Mian IS, Sjeander K, Haussler D (1994) J Mol Biol 235:1501–1531
14. Di Francesco V, Munson PJ, Garnier J (1999) Bioinformatics 15:131–140
15. James AJ (1995) Solving the many electron problem with quantum Monte-Carlo methods. Imperial College of Science, Technology and Medicine, London, pp 12–202
16. Landau LD, Lifshitz EM (1963) Mecánica Quántica no-Relativista. In: Curso de Física Teórica, vol 3. Reverté, Barcelona, pp 1–49
17. Dreizler RM, Gross EKU (1990) Density functional theory: an approach to the quantum many-body problem. Springer, Berlin Heidelberg New York, pp 1–30
18. Estrada E, Uriarte E (2001) Curr Med Chem 8:1573–1588
19. Kubinyi H (1999) J Recept Signal Transduct Res 19:15–39
20. Kier LB, Hall LH (1999) Topological indices and related descriptors in QSAR and QSPR. Gordon and Breach, Amsterdam, pp 455–489
21. Devillers J, Balaban AT (2000) Topological indices and related descriptors in QSAR and drug design. Amsterdam, pp 3–41
22. Hall LH, Kier LB (1977) Tetrahedron 33:1953–1957
23. Bonchev D (1983) Information theoretic indices for characterization of chemical structure. RSP-Wiley, Chichester, UK, pp 4–20
24. Trinjastic N (1992) Chemical graph theory. CRC Press, Boca Raton, Fla., pp 1–30
25. Bonchev D, Rouvray DH (1991) Chemical graph theory. Gordon and Breach, New York, pp 6–23
26. Dewar MJ (1991) MOTEC 91, modern technique in computational chemistry. Leiden, pp 6–15
27. Ögnetir C, Csizmadia IG (1991) Computational advances in organic chemistry: molecular structure and reactivity. Kluwer, Dordrecht
28. Kikuchi O (1987) Quant Struct–Act Relat 6:179–184
29. Gajewski JJ, Gilbert KE, Mckelvey J (1990) Advances in molecular modeling. JAI Press, Greenwich, pp 65–68
30. Estrada E (1997) J Chem Inf Comput Sci 37:320–328
31. Estrada E (1996) J Chem Inf Comput Sci 36:844–849
32. Stewart JJ (1989) J Comput Chem 10:209–221
33. Dewar MJ, Zoebish EG, Healy EF, Stewart JJ (1985) J Am Chem Soc 107:3902–3909
34. Todeschini R, Consonni V (2000) Handbook of molecular descriptors. Wiley-VCH, Weinheim
35. Denny WA (1992) The role of medicinal chemistry in the discovery of DNA-active anticancer drugs: from random searching, through lead development, to the novo design. In: Waring MJ, Ponder BAJ (eds) The search for anticancer drugs. Kluwer, Dordrecht, chapter 2
36. Lunney EA (1998) Med Chem Res 8:352–361
37. Walters WP, Stahl MT, Murcko MA (1998) Drug Discovery Today 3:160–178
38. Kubinyi H (1999) J Recept Signal Transduct Res 19:15–39
39. Lien EJ, Lien LL (1998) Chin Phar J 50:249–256
40. Lunney EA (1998) Med Chem Res 8:352–361
41. Walters WP, Stahl MT, Murcko MA (1998) Drug Discovery Today 3:160–178
42. Drie JHV, Lajines MS (1998) Drug Discovery Today 3:274–283
43. Ferrante K, Winograd B, Canetta R (1999) Cancer Chemother Pharmacol 43:S61-S83
44. Menta E, Palumbo M (1998) Expert Opin Ther Pat 8:1627–1672
45. Estrada E, Gutiérrez Y, González DH (2000) J Chem Inf Comput Sci 40:1386–1399
46. Estrada E, Gonzalez DH (2003) J Chem Inf Comput Sci 43:75–84
47. Randič M (1991) J Math Chem 7:155–168
48. Padron JA, Carrasco R, Pellón RF (2002) J Pharm Pharmaceut Sci 5:267–274

49. González DH, Olazábal E, Castañedo N, Hernádez SI, Morales A, Serrano HS, González J, Ramos de Armas R (2002) J Mol Mod 8:237–245
50. González DH, Hernádez SI, Uriarte E, Santana L (2003) Comput Chem (in press, corrected proofs published online)
51. González DH, De Armas RR, Uriarte E (2002) Online J Bioinformatics 1:83–95
52. Cabrera MA, González DH, Teruel C, Pla-Delfina JM, Bermejo del Val M (2002) Eur J Pharm Biopharm 53:317–325
53. Randič M (1991) Chemom Intell Lab Sist 10:213–227
54. Hall LH, Mohney B, Kier LB (1991) Quant Struct–Act Relat 10:43–51
55. Gálvez J, García R, Salabert MT, Soler R (1994) J Chem Inf Comput Sci 34:520–525
56. Gnedenko B (1978) The theory of probability. Mir, Moscow, pp 107–112
57. Pauling L (1939) The nature of the chemical bond. Cornell University Press, Ithaca, N.Y., pp 2–60
58. Grimmett GR, Stirzaker DR (1992) Probability and random processes. Clarendon Press, Oxford, pp 194–264
59. Kier LB, Hall LH (1999) Molecular structure description. The electrotopological state. Academic Press, New York
60. Estrada E, Molina E (2001) J Chem Inf Comput Sci 41:791-797
61. Hernández I, González H (2002) MARCH-INSIDE version 1.0 (Markovian chemicals "in silico" design). Chemicals Bioactives Center, Central University of Las Villas, Cuba. This is a preliminary experimental version future professional version shall be available to the public. For any information about it, send an e-mail to the corresponding author humbertogd@cbq.uclv.edu.cu
62. Rogers KS, Camnarata A (1969) J Med Chem 12:692–693
63. Jiang Y, Tang A, Hoffman R (1984) Theor Chim Acta 66:183–192
64. Burdett JK, Lee S (1985) J Am Chem Soc 107:3063–3082
65. Burdett JK, Lee S (1985) J Am Chem Soc 107:3050–3063
66. Lee S (1991) Acc Chem Res 24:249–254
67. Markovick S, Gutman I (1991) J Mol Struct (THEOCHEM) 81:81–87
68. Gutman I, Rosenfield VR (1996) Theor Chim Acta 93:191–197
69. Gutman I (1992) Theor Chim Acta 83:313–318
70. Karwowski J, Bielinska-Waz D, Jurkowski J (1996) Int Quantum Chem 60:185–193
71. Estrada E, Peña A, García-Domenech R (1998) J Comp Aided Mol Design 12:583–595
72. STATISTICA for Windows (2001) release 6.0. Statsoft Inc
73. van Waterbeemd H (1995) Discriminant analysis for activity prediction. In: Manhnhold R, Krogsgaard-Larsen P, Timmerman H (eds) Method and principles in medicinal chemistry, vol 2. Chemometric methods in molecular design, van Waterbeemd H (ed). VCH, Weinhiem, pp 265–282
74. Julián-Ortiz JV, Gálvez J, Mullños-Collado C, García-Domenech R, Gimeno-Cardona C (1999) J Med Chem 42:3308–3314
75. Kowalski RB, Wold S (1982) Pattern recognition in chemistry. In: Krishnaiah PR, Kanal LN (eds) Handbook of statistics. North-Holland, Amsterdam, pp 673–697
76. Mc Farland JW, Cooper CB, Newcomb DM (1991) J Med Chem 34:1908–1911
77. Negwer M (1987) Organic chemical drugs and their synonyms. Akademie-Verlag, Berlin
78. Mc Farland JW, Gans DJ (1995) Cluster significance analysis. In: Manhnhold R, Krogsgaard-Larsen P, Timmerman H (eds) Method and principles in medicinal chemistry, vol 2. Chemometric methods in molecular design, van Waterbeemd H (ed). VCH, Weinhiem, pp 295–307
79. Johnson RA, Wichern DW (1988) Applied multivariate statistical analysis. Prentice-Hall, N.J.
80. Via DL, Gia O, Magno MS, Da Settimo A, Primofiore G, Da Settimo F, Simorini F, Marini AM (2002) Eur J Med Chem 37:475–486
81. Galvez J, Garcia-Domenech R, Gomez-Lechon MJ, astell JV (1996) Bioorg Med Chem Lett 6:2301–2306
82. Via DL, Gia O, Viola G, Bertoloni G, Santana L, Uriarte E (1998) Il Farmaco 53:638–644
83. Gia O, Anselmo A, Conconi MT, Antonello C, Uriarte E, Caffieri S (1996) J Med Chem 39:4489–4496
84. Via DL, Gia O, Magno MS, Santana L, Teijeira M, Uriarte E (1999) J Med Chem 42:4405–4413
85. Fejzo J, Lepre ChA, Peng WJ, Bemis WG, Ajay MAM, Moore MJ (1999) Chem Biol 6:755–769
86. Gramatica P, Corradi M, Consonni V (2000) Chemosphere 41:763–777
87. Estrada E, Uriarte E, Montero A, Teijeira M, Santana L, De Clercq E (2000) J Med Chem 43:163–166
88. Frank IE, Todeschini R (1994) The data analysis handbook. Elsevier, Amsterdam
89. Gia O, Uriarte E, Zagotto G, Baccichetti F, Antonello C, Marciani-Magno S (1992) J Photochem Photobiol B: Biol 14:95–104
90. Gia O, Via LD, Marciani S, Angelini G, Margonelli A, Rodighiero P (2000) Photochem Photobiol B 56:132–138
91. Rodighiero P, Pastorini G, Via LD, Gia O, Marciani S (1998) Il Farmaco 53:313–319
92. Hermann T, Westhof E (2000) Combinatorial Chem High Throughput Screening 3:219–234
93. Gozalbes R, Gálvez J, García-Domenech R, Derouin F (1999) SAR QSAR Environ Res 10:47–60
94. Foye WO, Lemke TL, Williams DA (1995) Principles of medicinal chemistry. Williams and Wilkins, Baltimore, Md., pp 896–900
95. Arabzadeh A, Bathaiee SZ, Farsam H, Amanlou M, Saboury AA, Shockravi A, Moosavi-Movahedi AA (2002) Int J Pharmaceutics 237:47–45
96. Spicer JA, Swarna GA, Graene JF, Denny FW (2002) Bioorg Med Chem 10:19–29
97. Dollery C, Boobis A, Rawlins M, Thomas S, Wilkins M (1999) Therapeitic drugs. Churchill Livingston, Edimburgh, pp 102–108
98. Scott BR, Pathak MA, Mohn GR (1976) Mutat Res 39:29–74
99. Bridges BA, Mottershead RP (1977) Mutat Res 44:305–312
100. Ekins S, Boulanger B, Swaan WP, Hucpey AZ (2002) J Comput Aided Mol Des 16:381-401