Bioorganic & Medicinal Chemistry 19 (2011) 6239–6244

Contents lists available at SciVerse ScienceDirect

Bioorganic & Medicinal Chemistry

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

Multi-target drug discovery in anti-cancer therapy: Fragment-based approach toward the design of potent and versatile anti-prostate cancer agents

Alejandro Speck-Planche [a,*], Valeria V. Kleandrova [b], Feng Luan [a], M. Natália D.S. Cordeiro [a,*]

[a] REQUIMTE/Department of Chemistry and Biochemistry, University of Porto, 4169-007 Porto, Portugal
[b] Faculty of Technology and Production Management, Moscow State University of Food Production, Volokolamskoe Shosse 11, Moscow, Russia

ARTICLE INFO

Article history:
Received 15 May 2011
Revised 24 July 2011
Accepted 8 September 2011
Available online 13 September 2011

Keywords:
Multi-target QSAR
Prostate cancer
Linear discriminant analysis
Fragments
Quantitative contributions

ABSTRACT

Prostate cancer (PCa) is the second-leading cause of cancer deaths among men in the around the world. Understanding the biology of PCa is essential to the development of novel therapeutic strategies, in order to prevent this disease. However, after PCa make metastases, chemotherapy plays an extremely important role. With the pass of the time, PCa cell lines become resistant to the current anti-PCa drugs. For this reason, there is a necessity to develop new anti-PCa agents with the ability to be active against several PCa cell lines. The present work is an effort to overcome this problem. We introduce here the first multi-target approach for the design and prediction of anti-PCa agents against several cell lines. Here, a fragment-based QSAR model was developed. The model had a sensitivity of 88.36% and specificity 89.81% in training series. Also, the model showed 94.06% and 92.92% for sensitivity and specificity, respectively. Some fragments were extracted from the molecules and their contributions to anti-PCa activity were calculated. Several fragments were identified as potential substructural features responsible of anti-PCa activity and new molecular entities designed from fragments with positive contributions were suggested as possible anti-PCa agents.

1. Introduction

Prostate cancer (PCa) is the most enigmatic of the common solid malignancies. Second only to lung cancer as a killer of men beyond middle age, it warrants more attention than it currently receives from governments, researchers and the general public worldwide.[1] With the exception of skin cancers, PCa is the most commonly diagnosed cancer among men and the second leading cause of cancer death in many industrialized nations. In 2004, the American Cancer Society estimates that 29,900 American men will die from PCa and 230,110 will be diagnosed with this disease.[2] New technologies such as the da Vinci robot to facilitate laparoscopic radical prostatectomy and low dose brachytherapy both offer this prospect, and are rapidly becoming the dominant active treatment options in regions such as North America.[1] By contrast, locally advanced PCa is probably best managed by conformal external beam radiotherapy (EBRT) with pre and sometimes post-treatment hormonal ablation, although high dose brachytherapy is looking very interesting. However, once prostate cancer has spread to either lymph nodes or bones, hormonal therapy is usually the first line of treatment and may be effective for many months or years. Eventually, however, hormone relapse develops and second line treatments need to be considered. Chemotherapy with taxotere has now been shown to improve survival and several newer therapies in this context including endothelin A antagonists look promising.[3] In this sense, the search of new and more potent anti-PCa agents constitutes a challenge for the scientific community in order to avoid problems related with resistance to the current anti-PCa drugs.

In the last 10 years, several chemical families of compounds have been synthesized and evaluated for anti-PCa activity and computer-aided drug design (CADD) methodologies have played an important role toward the design of anti-PCa compounds.[4–10] Some disadvantages of the CADD methodologies employed in drug discovery for anti-PCa therapy is that they have been applied to small and homogeneous databases of compounds usually considering only one PCa cell line or referred to only one target like protein associated with PCa. In this sense, the discovery of anti-PCa compounds against several important and highly cited PCa cell lines constitutes a major interest. In order, to overcome this problem we develop the first multi-target approach for the design and prediction of anti-PCa agents against several cell lines. This approach represents a fragment-based QSAR model using a heterogeneous database of compounds for the efficient and fast extraction of substructural alerts responsible of anti-PCa activity depending on the type of PCa cell line.

* Corresponding authors. Fax: +351 220402659 (A.S.P.); fax: +351 220402659 (M.N.D.S.C.).
E-mail addresses: alejspivanovich@gmail.com (A. Speck-Planche), ncordeir@fc.up.pt (M. Natália D.S. Cordeiro).

0968-0896/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2011.09.015

2. Materials and methods

2.1. Functional group counts

Functional group counts (FGCs) are descriptors which express certain fragmental features. They are simple molecular descriptors defined as the number of specific functional groups in a molecule and they are calculated from the molecular composition and atom connectivity.[11–13] Many of the functional groups defined here, are those which are traditionally used in Organic Chemistry. FGCs are descriptors that have relation with the indicator variables in a Free-Wilson analysis.[14]

2.2. Atom-centered fragments

Atom-centered fragments (ACFs) have demonstrated to be very useful descriptors, and have been employed in some QSAR studies.[15–18] As the FGCs descriptors, ACFs are related with variable of the Free-Wilson analysis. They provide important information about hydrophobic and dispersive interactions which are involved in biological processes such as transport and distribution of drugs through the membrane[19] and about drug–receptor interactions.[17,19–21] ACFs are simple molecular descriptors which are defined as the number of specific atom types in a molecule. They are calculated from the molecular composition and atom connectivities. Each type of atom in the molecule is described in terms of its neighboring atoms. Hydrogen and halogen atoms are classified by the hybridization and oxidation states of the carbon atom to which they are attached. For hydrogen atoms, heteroatoms which are attached to a carbon in α-position are further considered. Carbon atoms are classified by their hybridization state and depending on whether their neighbors are carbon or heteroatoms.

2.3. Spectral moments of the bond adjacency matrix

The approach that encloses the calculation of the spectral moments of the bond adjacency matrix, is known as TOPS-MODE (TOPological Substructural MOlecular DEsign) approach and it has been applied for the modeling of pharmacological activities,[22–24] and in quantitative structure-toxicity relationship (QSTR) studies.[25–29] The theoretical background about the spectral moments of bond adjacency matrix has been described in many papers; however, we will focus our explanation on the most important aspects. In this approach the molecular structure is encoded by mean of the edge adjacency matrix **E** (commonly called the bond adjacency matrix **B**). The **E** or **B** matrix is a square table of order m (the number of chemical bonds in the molecule). The elements of this matrix (e_{ij}) are equal to 1 if bonds i and j are adjacent (which means that i and j are incident in the same vertex or atom) and 0 otherwise. In order to encode information of heteroatoms, the TOPS-MODE approach uses $\mathbf{B}(w_{ij})$ weighted matrices instead of **B**. The weights (w_{ij}) are chemically meaningful numbers such as bond distances, bond dipoles, or mathematical expressions involving atomic weights. The weights are introduced in the main diagonal of matrix $\mathbf{B}(w_{ij})$. Then, the spectral moments of this matrix can be used as molecular fingerprints in QSAR studies for the codification of molecular structures. By mathematical definition, the term spectral moments must be understood as the sum of the elements (e_{ij}) in the natural powers of $\mathbf{B}(w_{ij})$.[30–32] Then the spectral moment of order k ($\mu_{\mathbf{k}}$) is the sum of the main diagonal elements (e_{ii}) of matrix $\mathbf{B}(w_{ij})^k$. The spectral moments of the bond matrix are defined as:

$$\mu_k = Tr(E^k) = \sum_{i=j} (e_{ij})_k \quad (1)$$

where **Tr** means the trace of the matrix, that is the sum of the diagonal entries of the matrix and the elements $(e_{ii})_k$ are the diagonal entries of the kth power of the bond matrix. The spectral moments of the bond adjacency matrix have topological nature. The principal advantage of these descriptors is the possibility to calculate the relative contribution of any fragment to the desired activity.[33] That is possible because they can be expressed as linear combinations of the number of times in which a fragment appears in the molecule. Another advantage of the spectral moments of the bond adjacency matrix is the ability to explain in a reasonable way, a considerable part of spatial phenomena.[34] That is a particular characteristic of topographical descriptors.

2.4. Selection of the data set: calculation of the descriptors and development of the model

The data set was formed by 816 compounds with anti-PCa activity against four PCa cell lines.[35] These PCa cell lines were: **CWR22R**, **LAPC4**, **LNCaP** and **PC-3**. Not all the compounds were tested against all the subtypes. We had also 213 drugs which have been reported in the Merck Index. These compounds present other activities that do not include anti-PCa activity and have been used as inactive.[36] The FGCs and ACFs were calculated using DRAGON v5.3.[11] The spectral moments of the weighted bond adjacency matrix (from order 1 to 15), were calculated using MODESLAB v1.5.[37] In this case, the spectral moments were weighted by physicochemical properties such as atomic weights and by two Abraham terms: the first containing information about the relation dipolarity/polarizability and the second containing information about the hydrogen bond basicity. Linear Discriminant Analysis (LDA) was used to construct the classifier model.[38] The most important step in this work was the organization of the spreadsheet containing the raw data used as input for the LDA, because this was not a classical LDA problem. For this reason we employed the multi-target QSAR methodology which was applied by González-Díaz and coworkers for the prediction of enzyme classes in *Leishmania infantum*.[39] Here, we formulated a 2 group discriminant function to classify compounds: compounds that belonged to a particular group (anti-PCa activity against a specific cell line) and compounds that did not belong to this group (inactive). For this, we had to construct an approach to define the groups predicted in each case. The following steps were used:

First, the 816 compounds belonging to the group of compounds with anti-PCa activity were divided according to their activity against different PCa cell lines. Each PCa cell line had a cutoff value of anti-PCa activity. This value represented that at which a compound was considered as active. The variable of activity selected was IC_{50}, the concentration that inhibits at 50% the proliferation (or growth) of the PCa cell line. So, the cutoff values of IC_{50} to consider the compounds as active were the following: $\leqslant$5.0 μM for **CWR22R**, $\leqslant$6.0 μM for **LAPC4**, $\leqslant$10.0 μM for **LNCaP** and $\leqslant$9.8 μM for **PC-3**.

We created a raw data file by assigning to each compound 322 structural variables (inputs): 99 FGCs, 88 ACFs and 135 descriptors like μ_k. We had also; one output variable and one classification variable related the type of PCa cell line (PCaCL). This last variable is an auxiliary variable which was not used to construct the model. Thus, the row for each compound input contained 324 elements in total.

The output variable is a dummy variable (Boolean) called anti-PCa activity (A_{PCa}); $A_{PCa} = 1$ if the compound has anti-PCa activity against a defined type of PCa cell line and -1, otherwise ($A_{PCa} = -1$). The last case corresponds to the 213 drugs which were considered as inactive. We can repeat each of these 213 drugs more than one time in the raw data. In fact, we repeated

each compound four times corresponding to the four PCa cell lines. In this work, we used only the PCaCL code which relates a compound with its corresponding type of PCa cell line, and thus, these compounds entered only once. Conversely, inactive compounds (decoys) have more than one line entry with different PCaCL classes.

The problem in this type of organization for raw data is that the 'traditional' 45 μ_k are not able to discriminate the structural information about all the anti-PCa compounds which are active against the different PCa cell lines. Consequently, a LDA model based only on 45 μ_k (or any of the other 99 FGCs or 88 ACFs) will necessarily fail. The problem is that we need four specific probabilities: each of them confirming the real PCaCL class. We can solve this problem by introducing variables characteristic of each PCaCL class. For this, we used the average value of each spectral moment ($\bar{\mu}_k$) for all compounds that belong to the same PCaCL class. We also calculated the deviation of the μ_k from the respective group ($\Delta\mu_k$) indicated in PCaCL. In the case of these last descriptors they were calculated as the difference between the original descriptor of each compound and $\bar{\mu}_k$. In total, we therefore have (45 μ_k values) + (45 $\bar{\mu}_k$ values) + (45 $\Delta\mu_k$ like deviation values) = 135 input variables like spectral moments. It is very important to understand that we never used PCaCL as input, so the model only include as input FGCs, ACFs, the μ_k values for the compounds entry and $\bar{\mu}_k$ and $\Delta\mu_k$ values from the PCaCL. The names or codes of all the compounds used in this work appear in the supplementary material 1 file (Supplementary data 1).

We were able to collect 1668 cases (compounds/PCaCL pairs) in total. In order to perform a rigorous design, the dataset was divided in two series: training and prediction series. The training set was formed by 1250 cases, 610 of them considered anti-PCa agents and 640 inactive compounds, and the prediction set was composed by 418 cases, 206 with anti-PCa activity and 212 inactive compounds. The general expression for this LDA model is presented in the following form:

$$A_{PCa} = a_0 + \sum_k b_k \cdot x_k + \sum_k c_k \cdot \bar{x}_k + \sum_k d_k \cdot \Delta x_k \quad (2)$$

where A_{PCa} is not the probability, but a real value score that predicts the propensity of a compound to have anti-PCa activity. The term described as a_0 is the constant, b_k, c_k and d_k are the corresponding coefficients of the variables in the model. The symbol x_k represents the different (FGCs, ACFs and/or μ_k) descriptors while $\bar{x}_k$ and Δx_k are average and deviation values (referred only to μ_k), respectively. The discriminant function was obtained by employing the LDA modules of STATISTICA 6.0.[40] The default parameters of this program were used for the development of the model. The variables which were included in the discriminant model were selected using a forward stepwise procedure as the variable selection strategy. The selection of the best model was subjected also, to the principle of parsimony.[14] Then, the model with high statistical significance, but having as few parameters as possible, was chosen.

The statistical quality of the model was determined, examining some statistical indices such as the Wilks' lambda (λ), the chi-square χ^2 and the p-level and the proportion between compounds and variables used to develop the model ρ. Another important aspect is that the compounds used in the prediction set, were never used to develop the discrimination function. On the other hand, to confirm the quality of the model, and to validate it, we employed other statistics such as sensitivity (*sens*) as the ability for the classification of active cases, specificity (*spec*) as the ability for the classification of inactive cases and accuracy (*acc*) as the overall predictability. These statistical indices were calculated for both, training and prediction series according to the following equations:

$$sens = \frac{TP}{TA} \cdot 100\% \quad (3)$$

$$spec = \frac{TN}{TI} \cdot 100\% \quad (4)$$

$$acc = \frac{TP + TN}{TA + TI} \cdot 100\% \quad (5)$$

where **TP** means the cases (compounds) classified correctly by the model as active, **TA** the total active compounds, **TN** means the cases classified correctly by the model as inactive and **TI** represents the total inactive compounds.

2.4.1. ROC curve

The sensitivity and the specificity can describe adequately the quality of a model. However, these two statistical indices have disadvantages. The most important one is that they cannot provide information about how many times the probabilities indicate that a compound, observation or case will be predicted more as positive (active) than negative (inactive), and this is very important since it confirms together with the positive predictive value if a given case is active. However, that information can be provided by a Receiver-Operating Characteristic (ROC) analysis. ROC is a classic methodology from signal detection theory.[41] The ROC curve is created by plotting the true-positive rate against false-positive rate, or sensitivity against (1—specificity). The ROC curve going along the diagonal from bottom left to upper right represents pure-chance performance. Thus the area under the ROC curve can contribute in very important way to the assessment of the quality and predictive ability of a model as classifier.

3. Results and discussion

3.1. Discriminant model

Taking into consideration the previous ideas about the strategy of variable selection and the principle of parsimony, the best model obtained by us, contains 14 descriptors. This model had three types of descriptors calculated by us: FGCs, ACFs and descriptors like μ_k:

Table 1
Molecular descriptors used in the model

Descriptor	Definition[a]
R=Cs	Number of aliphatic secondary Csp_2 atoms
Ct	Number of total tertiary Csp_3
RCOOH	Number of aliphatic carboxyl groups
RCOOR	Number of aliphatic ester groups
$RCONR_2$	Number of disubstituted aliphatic amides
$ArNO_2$	Number of aromatic nitro groups
Pyrr	Presence or absence of pyrrole rings
H-046	Number of hydrogen atoms which are attached to C^0sp_3 no X attached to adjacent C
Cl-089	Number of Cl atoms attached to C^1sp_2 atoms
N-075	Number of Nsp_2 atoms present in the fragments of structures Z–N—Z*/Z–N–X
Br-094	Presence or absence of Br atoms attached to C^1sp_2 atoms
$\bar{\mu}_{15}^{(Ato)}$	Average spectral moment of order 15 weighted by the atomic weights
$\Delta\mu_{12}^{(Ab\pi 2H)}$	Deviation spectral moment of order 12 weighted by the Abraham dipolarity/polarizability term
$\Delta\mu_{12}^{(Ab\text{-}sum2H)}$	Deviation spectral moment of order 2 weighted by the Abraham hydrogen bonding basicity

* Pyridine-type structure.

[a] The superscript over carbon atoms represents the formal oxidation number. The formal oxidation number of a carbon atom equals the sum of the conventional bond orders with electronegative atoms; the C–N bond order in pyridine may be considered as 2 while we have one such bond and 1.5 when we have two such bonds; the C···X bond order in pyrrole or furan may be considered as 1. X represents any electronegative atom (O, N, S, P, Se, halogens). Z represents any group linked through carbon.

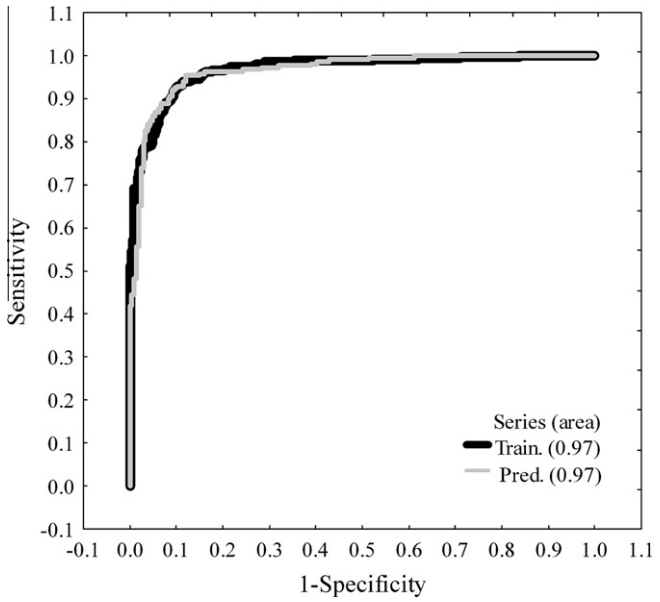

Figure 1. ROC curve.

$$A_{PCa} = 0.303(R = Cs) - 0.337(Ct) - 0.923(RCOOH) + 0.538(RCOOR) - 0.648(RCONR_2) + 1.013(ArNO_2) + 0.661(Pyrr) + 0.057(H - 046) - 0.515(CI - 089) + 0.622(N - 075) + 1.029(Br - 094) + 8.865 \cdot 10^{-30}\bar{\mu}_{15}^{(Ato)} - 4.291 \cdot 10^{-8}\Delta\mu_{12}^{Ab-\pi 2H} + 0.017\Delta\mu_{2}^{(Ab\text{-}SumB2H)} - 1.994$$

$$N = 1250\ \lambda = 0.347\ \chi^2 = 1312.15\ p < 0.001\ \rho = 89.28 \quad (6)$$

The symbology of the different descriptors together in the Eq. 6, with their corresponding meaning appear summarized in Table 1. It is necessary to point out that with 15 descriptors we raised a good performance of the model. The increment in the number of variables did not improve the quality of the model in a significant way, while a diminution in the number of variables, conducted to an appreciable lack of quality and predictive ability. The sensitivity of the model was 88.36% and the specificity 94.06% in the training series, for an accuracy of 91.28%. We examined all the compounds, searching misclassified cases because they can be outliers and this fact can have influence in the quality of a model. We checked the Mahalanobis distance of each molecule with respect the two centroids of both groups (active and inactive compounds). Generally, in the case of abnormal values, the case should be excluded of the model. In our case, no outliers were detected and the deletion of the misclassified compounds did not improve the quality of the model. In order to validate our model, we took into consideration the sensitivity, the specificity and the accuracy in the prediction series. The sensitivity of the model in prediction series was 89.81% and the specificity was 92.92%, for an accuracy of 91.39%. The names or codes, and probabilities of anti-PCa activity of each compound (expressed as percentages) are recorded in the supplementary material 2 file (Supplementary data 2). We took also as determinant statistics for the good performance of the model the areas under the ROC curves. These values were 0.97 for both, training and prediction series (Fig 1). These values of area can be interpreted as follows: in that value of area (0.97), means that a randomly selected compound or case from the active group will has a larger value of probability than a randomly selected compound or case from the inactive group, 97% of the times. A similar deduction can be made from the area under the ROC curve for prediction series. Altogether, this proves that our model is not a random classifier because the areas under the ROC curves are different and statistically significant from those obtained by random classifiers (area = 0.5). According to the statistical indices and the values of sensitivity, specificity, accuracy and areas under the ROC curves, the model has a good quality and this in strong agreement with the reports of the literature.[39,42–46]

3.2. Extracting substructural alerts responsible of anti-PCa activity

It should be understood that the model obtained by the present approach has two principal advantages. First, the model can determine quickly and efficiently the probability of a compound to be an anti-PCa agent taking into consideration the different PCa cell lines.

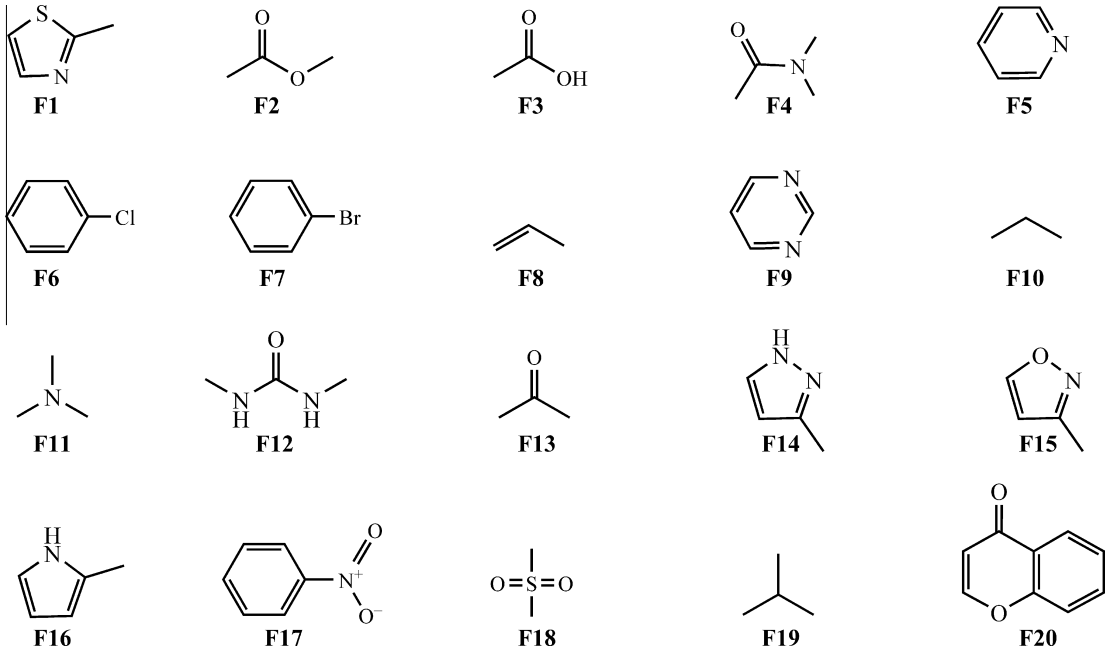

Figure 2. Different fragments which were found in the molecules.

Table 2
Quantitative contributions of some fragments to anti-PCa activity

ID	CWR22R	LAPC4	LNCaP	PC-3
F1	0.019	0.097	1.265	1.018
F2	−0.167	−0.089	1.079	0.833
F3	−1.662	−1.584	−0.416	−0.663
F4	−1.288	−1.210	−0.041	−0.288
F5	−0.014	0.064	1.233	0.986
F6	−1.084	−1.006	0.162	−0.085
F7	0.459	0.538	1.706	1.459
F8	−0.333	−0.255	0.913	0.666
F9	0.609	0.687	1.856	1.609
F10	−0.468	−0.390	0.778	0.531
F11	−0.738	−0.660	0.508	0.261
F12	−0.666	−0.588	0.581	0.334
F13	−0.739	−0.661	0.507	0.260
F14	0.021	0.100	1.268	1.021
F15	0.019	0.097	1.265	1.018
F16	0.059	0.137	1.306	1.059
F17	0.543	0.621	1.790	1.543
F18	−0.741	−0.663	0.505	0.258
F19	−0.565	−0.487	0.682	0.435
F20	−0.014	0.064	1.232	0.985

Table 3
Probabilities of the suggested molecules to have anti-PCa activity

ID	PCa cell lines			
	CWR22R	LAPC4	LNCaP	PC-3
M1	96.06	96.79	99.87	99.74
M2	93.73	94.88	99.78	99.57
M3	93.73	94.88	99.78	99.57
M4	93.78	94.92	99.78	99.57
M5	94.34	95.38	99.80	99.61
M6	99.70	99.76	99.99	99.98

The second advantage of the model is related to the nature and interpretation of the descriptors. Some of them express specific contributions of characteristic fragments which are present in the compounds with anti-PCa activity. This is the case of the FGCs and ACFs. These descriptors will contain structural information, essentially describing hydrophobic interactions of specific fragments with the biological receptors it was stated in previous works.[15–21] The signs of the coefficients in the equation will express the influence (favorable or unfavorable) of corresponding fragments to anti-PCa activity. On the other hand, descriptors like spectral moments take into account some physicochemical properties. For example, descriptor such as $\bar{\mu}_{15}^{(Ato)}$ encodes information related to the increment of molecular accessibility depending on the type of PCa cell line against the compounds were tested. On the other hand, $\Delta\mu_{12}^{(Ab-\pi 2H)}$ encodes the ability of the compounds to interact with their respective biological targets (inside the PCa cell line) by using regions in which there is a diminution in the relation dipolarity/polarizability. This means that only regions in which dipole of the bonds decrease and the polarizability is increased will have positive influence in the development of the anti-PCa activity.

Finally $\Delta\mu_{2}^{(Ab-sumB2H)}$ gives information about the ability of the molecules to interact as hydrogen bonding acceptors. The increment of these regions will be favorable for the increment of the anti-PCa activity. The last two descriptors will depend on both: the structure of the molecules and the PCa cell line against which they were tested. As all the descriptors employed in the model encode fragments and at the same time comply with the rule of linear additivity, we can calculate the contribution of any fragment to anti-PCa activity. These molecular fragment contributions can indicate the potential relation between molecular fragments with the anti-proliferative activity against different PCa cell lines. Thus, some fragments were represented (Fig. 2) and their contributions to the anti-proliferative activity against four different PCa cell lines were calculated and summarized in Table 2. If we compare the variation in the contribution of each fragment to anti-PCa activity in the four cell lines we can deduce that **CWR22R** and **LAPC4** could be less sensitive to the anti-PCa compounds than **LNCaP** and **PC-3**. On the other hand, some suitable fragments can be chosen for the design of new anti-PCa agents. For example, **F7**, **F9** and **F17** (in less degree, **F1** and **F14-F16**) can be considered as potential fragments for the discovery of anti-PCa agents because they have appreciable high positive contribution against the four PCa cell lines. Thus, in principle, new molecular entities can be generated from these fragments. We need to point out that some fragments with positive contribution can be in inactive molecules. At the same time, some fragments with negative contribution can be in molecules with anti-PCa activity. It is evident that only the combination of the different fragments will determine if a given molecule will be anti-PCa agent or not. Also, another important aspect of the model is related with the information provided by the descriptors, because

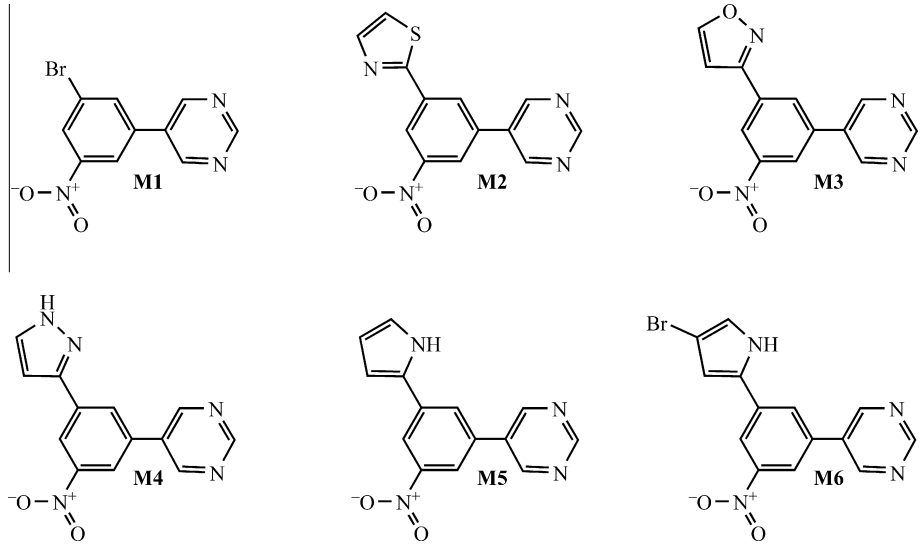

Figure 3. New molecular entities suggested as possible anti-PCa agents.

they are sensitive to small structural variations in the molecules, and for this reason they can result very useful in the design of anti-PCa agents.

3.3. New molecular entities as possible anti-PCa agents

In an attempt to show how our approach works, we extracted the fragments represented in the Figure 2 to construct new molecular entities. In this sense we designed 6 molecules taking into consideration the favorable contributions of the fragments mentioned above to the anti-PCa activity (Fig. 3). Thus we were able to calculate the probabilities of these molecules to anti-PCa agents against the four PCa cell lines. All the molecules were predicted by the model where their probabilities to be anti-PCa agents were calculated in order to check if the design was correctly realized. The probabilities of the molecules are higher than 90% (Table 3). These compounds could be, in principle, synthesized and after, they could be tested against the four PCa cell lines.

4. Conclusions

In this work a multi-target QSAR model based on substructural descriptors and using a large heterogeneous database of compounds was developed for the search, design and prediction of anti-PCa agents against different PCa cell lines. This fact can allows us to predict anti-PCa activity of compounds in more general situations than classical QSAR models which have as principal limitation the assessment of biological activity against only one type of PCa cell line. Our model has an appropriate statistical quality and as important element, it provides a promising methodology for the efficient and fast extraction of fragments which can be considered as substructural alerts responsible of the anti-PCa. Our model permitted also, to suggest new molecular entities that could be experimentally analyzed for the future assessment of the anti-PCa activity. To our knowledge, this work constitutes the first in silico methodology for the design of anti-PCa agents which may be able to inhibit more than one PCa cell line.

Acknowledgments

The authors acknowledge the Portuguese Fundação para a Ciência e a Tecnologia (FCT) and the European Social Found for financial support (Project PTDC/QUI-QUI/113687/2009 and Grant SFRH/BPD/63666/2009).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2011.09.015.

References and notes

1. Kirby, R. S.; Partin, A. W.; Parsons, J. K.; Feneley, M. R. *Treatment Methods for Early and Advanced Prostate Cancer*; Informa Healthcare: London, 2008.
2. Jemal, A.; Tiwari, R. C.; Murray, T.; Ghafoor, A.; Samuels, A.; Ward, E.; Feuer, E. J.; Thun, M. J. C. A. *Cancer J. Clin.* **2004**, *54*, 8.
3. Chiappori, A. A.; Haura, E.; Rodriguez, F. A.; Boulware, D.; Kapoor, R.; Neuger, A. M.; Lush, R.; Padilla, B.; Burton, M.; Williams, C.; Simon, G.; Antonia, S.; Sullivan, D. M.; Bepler, G. *Clin. Cancer Res.* **2008**, *14*, 1464.
4. Sarswat, A.; Kumar, R.; Kumar, L.; Lal, N.; Sharma, S.; Prabhakar, Y. S.; Pandey, S. K.; Lal, J.; Verma, V.; Jain, A.; Maikhuri, J. P.; Dalela, D.; Kirti; Gupta, G.; Sharma, V. L. *J. Med. Chem.* **2011**, *54*, 302.
5. Hassan, H. M.; Elnagar, A. Y.; Khanfar, M. A.; Sallam, A. A.; Mohammed, R.; Shaala, L. A.; Youssef, D. T.; Hifnawy, M. S.; El Sayed, K. A. *Eur. J. Med. Chem.* **2011**, *46*, 1122.
6. Khanfar, M. A.; El Sayed, K. A. *Eur. J. Med. Chem.* **2010**, *45*, 5397.
7. Mudit, M.; Khanfar, M.; Muralidharan, A.; Thomas, S.; Shah, G. V.; van Soest, R. W.; El Sayed, K. A. *Bioorg. Med. Chem.* **2009**, *17*, 1731.
8. Soderholm, A. A.; Viiliainen, J.; Lehtovuori, P. T.; Eskelinen, H.; Roell, D.; Baniahmad, A.; Nyronen, T. H. *J. Chem. Inf. Model.* **2008**, *48*, 1882.
9. Holder, S.; Lilly, M.; Brown, M. L. *Bioorg. Med. Chem.* **2007**, *15*, 6463.
10. Wang, D. F.; Wiest, O.; Helquist, P.; Lan-Hargest, H. Y.; Wiech, N. L. *Bioorg. Med. Chem. Lett.* **2004**, *14*, 707.
11. Talete-srl. *DRAGON for Windows (Software for Molecular Descriptor Calculations). v5.3*, 2005.
12. Todeschini, R.; Consonni, V. *Molecular Descriptors for Chemoinformatics*; WILEY-VCH Verlag GmbH & Co. KGaA: Weinheim, 2009.
13. Crowe, J. E.; Lynch, M. F.; Town, W. G. *J. Chem. Soc. C* **1970**, *23*, 990.
14. Kubinyi, H. *QSAR: Hansch analysis and related approaches*; VCH Publishers: Weinheim, New York, Basel, Cambridge, Tokyo, 1993.
15. Speck-Planche, A.; Scotti, M. T.; García-López, A.; Emerenciano, V. P.; Molina-Pérez, E.; Uriarte, E. *Mol. Divers.* **2009**, *13*, 445.
16. Speck-Planche, A.; Scotti, M. T.; Emerenciano, V. P.; García-López, A.; Molina-Pérez, E.; Uriarte, E. *J. Comput. Chem.* **2010**, *31*, 882.
17. Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163.
18. Speck-Planche, A.; Guilarte-Montero, L.; Yera-Bueno, R.; Rojas-Vargas, J. A.; Garcia-Lopez, A.; Uriarte, E.; Molina-Perez, E. *Pest. Manag. Sci.* **2011**, *67*, 438.
19. Ghose, A. K.; Crippen, G. M. *J. Comput. Chem.* **1986**, *7*, 565.
20. Viswanadhan, V. N.; Reddy, M. R.; Bacquet, R. J.; Erion, M. D. *J. Comput. Chem.* **1993**, *14*, 1019.
21. Ghose, A. K.; Pritchett, A.; Crippen, G. M. *J. Comput. Chem.* **1988**, *9*, 80.
22. Estrada, E.; Molina, E.; Nodarse, D.; Uriarte, E. *Curr. Pharm. Des.* **2010**, *16*, 2676.
23. Pisco, L.; Kordian, M.; Peseke, K.; Feist, H.; Michalik, D.; Estrada, E.; Carvalho, J.; Hamilton, G.; Rando, D.; Quincoces, J. *Eur. J. Med. Chem.* **2006**, *41*, 401.
24. Estrada, E.; Uriarte, E.; Molina, E.; Simon-Manso, Y.; Milne, G. W. *J. Chem. Inf. Model.* **2006**, *46*, 2709.
25. Estrada, E.; Uriarte, E. S. A. R. *QSAR Environ. Res.* **2001**, *12*, 309.
26. Estrada, E.; Uriarte, E.; Gutierrez, Y.; Gonzalez, H. S. A. R. *QSAR Environ. Res.* **2003**, *14*, 145.
27. Helguera, A. M.; Gonzalez, M. P.; Cordeiro, M. N. D. S.; Perez, M. A. *Toxicol. Appl. Pharmacol.* **2007**, *221*, 189.
28. Morales Helguera, A.; Perez Gonzalez, M.; Cordeiro, M. N. D. S.; Cabrera Perez, M. A. *Chem. Res. Toxicol.* **2008**, *21*, 633.
29. Helguera, A. M.; Perez-Machado, G.; Cordeiro, M. N. D. S.; Combes, R. D. S. A. R. *QSAR Environ. Res.* **2010**, *21*, 277.
30. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 844.
31. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 320.
32. Estrada, E. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 23.
33. Perez Gonzalez, M.; Gonzalez Diaz, H.; Molina Ruiz, R.; Cabrera, M. A.; Ramos de Armas, R. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1192.
34. Estrada, E.; Molina, E.; Perdomo-Lopez, I. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1015.
35. ChEMBL Database. http://www.ebi.ac.uk/chembldb/.
36. O'Neill, M. J.; Heckelman, P. E.; Koch, C. B.; Roman, K. J. *The Merck Index, An Encyclopedia of Chemicals, Drugs and Biologicals*; Whitehouse Station, NJ: Merck & Co., Inc: New Jersey, 2006.
37. Estrada, E.; Gutiérrez, Y. *MODESLAB v1.5*, 2002–2004.
38. van de Waterbeemd, H. *Chemometrics methods in molecular design*; VCH Publishers.: Weinheim, New York, Basel, Cambridge, Tokyo, 1995.
39. Concu, R.; Dea-Ayuela, M. A.; Perez-Montoto, L. G.; Bolas-Fernandez, F.; Prado-Prado, F. J.; Podda, G.; Uriarte, E.; Ubeira, F. M.; Gonzalez-Diaz, H. *J. Proteome. Res.* **2009**, *8*, 4372.
40. StatSoft. *STATISTICA v6.0. Data analysis software system* 2001.
41. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R. *Bioinformatics* **2010**, *26*, 822.
42. Gonzalez-Diaz, H.; Muino, L.; Anadon, A. M.; Romaris, F.; Prado-Prado, F. J.; Munteanu, C. R.; Dorado, J.; Sierra, A. P.; Mezo, M.; Gonzalez-Warleta, M.; Garate, T.; Ubeira, F. M. *Mol. Biosyst.* **2011**, *7*, 1938.
43. Marrero-Ponce, Y.; Khan, M. T.; Casanola-Martin, G. M.; Ather, A.; Sultankhodzhaev, M. N.; Garcia-Domenech, R.; Torrens, F.; Rotondo, R. *J. Comput. Aided. Mol. Des.* **2007**, *21*, 167.
44. Marrero-Ponce, Y.; Marrero, R. M.; Torrens, F.; Martinez, Y.; Bernal, M. G.; Zaldivar, V. R.; Castro, E. A.; Abalo, R. G. *J. Mol. Model.* **2006**, *12*, 255.
45. Speck-Planche, A.; Cordeiro, M. N. D. S. *Curr. Bioinform.* **2011**, *6*, 81.
46. Casañola-Martin, G. M.; Marrero-Ponce, Y.; Khan, M. T.; Khan, S. B.; Torrens, F.; Perez-Jimenez, F.; Rescigno, A.; Abad, C. *Chem. Biol. Drug Des.* **2010**, *76*, 538.