Quantitative structure–activity relationship (QSAR) modeling is one of the major computer aided modeling employed in medicinal chemistry. It is used for developing relationships between the effects (e.g. activities and properties of interest) of a series of molecules with their structural properties. The aim of this work was to develop QSAR models to predict the inhibition of $V600EBRAF$-dependent extracellular-regulated kinase (ERK) phosphorylation in WM266.4 melanoma cell lines ($IC_{50}$p ERK) using the CORAL software (http://www.insilico.eu/coral). These models make use of descriptors based on simplified molecular input-line entry system (SMILES), optimized with the Monte Carlo method. The statistical quality of the newly built models was satisfactory on three random splits of data.

Keywords: QSAR; Kinase inhibitors; Monte carlo method; Sorafenib; Smiles

Introduction

Mitogen-activated protein kinase (MAPK) modulates extracellular signals to control several cell functions (from proliferation and survival to differentiation and senescence)[1,2]. One of the most studied MAPK pathways is the extracellular signal-regulated kinase (ERK) pathway. ERK is a subgroup of MAPKs that is activated by external factors such as growth factors and mitogens[3]. The cascade of ERK signalling is triggered by the activation of receptor tyrosine kinases and flows through RAS GTPase[4]. RAF kinases are key components of this pathway: once RAS is turned on, it recruits and activates proteins necessary for the propagation of growth factor and other receptor signals, such as RAF. There are three RAF proteins in mammals, ARAF, BRAF, and CRAF, and they can all activate MAP kinase (MEK) just upstream of ERK[5]. The over-expression or mutations of the components of the RAS–RAF–MEK–ERK–MAP kinase pathway have been found to occur in up to 30 % of human cancers[6]. In different cancer types, the mutation of RAS or RAF causes the hyper activation of ERK followed by un-regulated cell proliferation[7], suggesting that inhibition of ERK represents a potential approach for the treatment of cancer[8]. In particular, BRAF somatic missense mutations have been found to occur in 66% of malignant melanomas and at lower frequency in a wide range of human cancers[5,9].

Among the BRAF mutations, a point mutation resulting in substitution of glutamic acid for valine, also known as BRAFV600E, is the most common change[6]. In the recent past, a number of small molecules inhibiting mutant RAF kinases have been discovered, and some of these have now advanced into clinical trials. For example, Sorafenib[10] (Figure 1) is the

Copyrights: © 2017 K, Bouhedjar. This is an Open access article distributed under the terms of Creative Commons Attribution 4.0 International License.
The most studied inhibitor for this class of kinases. It was approved by FDA in 2005 for the treatment of renal cell carcinoma and in 2007 for the treatment of hepatocellular carcinoma, and is still undergoing multiple clinical trials in other types of cancers[11,12].

The experimental measurement of the inhibition activity of chemicals is difficult, expensive and time-consuming. QSAR based analysis can be used as a tool to screen or filter anti-cancer drug candidates, before they are subjected to more intensive calculations, such as docking, or to experimental in vitro measurement of activity and finally under in vivo conditions[13]. In this study, a model was built to predict the half maximal inhibitory concentration (IC_{50}) of V600EBRAF-dependent ERK phosphorylation in WM266.4 cells. This model made use of descriptors based on simplified molecular input-line entry system (SMILES)[14-17], calculated with the CORAL software. Then, chemical information obtained from the calculation of these descriptors was used for drawing up hypothesis of molecular design of Sorafenib derivatives.

The aim was to obtain molecules able to inhibit 50% of ERK phosphorylation in WM266.4 melanoma cells at lower concentrations compared to the template molecule.

Method

Data

The dataset of 142 chemical structures used in this study was taken from the literature[18,19]. All these compounds shared a common scaffold (Figure 2). The dataset contained chemical structures represented as SMILES strings, and experimental values relative to the inhibition of the phosphorylation of extracellular signal-regulated kinases (p-ERK).

![Figure 2: General structure of target series.](image)

The selected endpoint was IC_{50} of V600EBRAF-dependent ERK phosphorylation in WM266.4 melanoma cells (IC_{50} pERK)[11,12]. IC_{50} values were expressed as negative decimal logarithm (pIC_{50}). In this study ACD/ChemSketch[21], was used to build up SMILES of each structure.

CORAL and Optimal descriptors

CORAL is freely available standalone application software for building up regression or classification QSAR models based on the Monte Carlo technique[22]. The calculations are repeated several times for various splits into training, calibration and external validation sets. Optimal descriptors for constructing QSAR models are based on graph or SMILES. The molecular graph includes hydrogen suppressed graph (HSG), hydrogen filled graph (HFG), and graph of atomic orbital’s (GAO).

The SMILES-based optimal descriptors include a group of local and one of global attributes. Sk, SSk, and SSSk are local descriptors represented by a sequence of atoms and bonds present in the SMILES string[23].

The SMILES “ClCCCBr”, Sk, SSk, and SSSk can be represented as follows:

ClCCCBr → Cl + C + C + C + Br ( Sk )
ClCCCBr → ClC + CC + CC + CBr ( SSk )
ClCCCBr → ClCC +CCC+ CCBr ( SSSk )

BOND, PAIR, NOSP and HALO are global SMILES attributes, involving the presence of the following specific items in the target SMILES:

- **BOND** refers to the presence/absence of double (=), triple (#), and stereo chemical (@) bonds;
- **PAIR** indicates the co-occurrence of two elements among the followings: F, Cl, Br, I, N, O, S, P, #, =, @;
- **NOSP** indicates the presence/absence of nitrogen, oxygen, sulfur, and phosphorus;
- **HALO** refers to the presence/absence of halogens.

A more detailed explanation of CORAL descriptors has been provided by[24]. All the SMILES-based local descriptors and BOND, NOSP and HALO global attributes were selected to build the regression models.

Table 1: The statistical characteristics of QSAR model of pIC_{50} p-ERK inhibitors.

<table>
<thead>
<tr>
<th>Set</th>
<th>n</th>
<th>r^2</th>
<th>q^2</th>
<th>s</th>
<th>MAE</th>
<th>F</th>
<th>cR^2_y</th>
<th>T</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Split 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>86</td>
<td>0.90</td>
<td>0.91</td>
<td>0.371</td>
<td>0.289</td>
<td>743</td>
<td>0.89</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Calibration</td>
<td>28</td>
<td>0.83</td>
<td>0.80</td>
<td>0.481</td>
<td>0.356</td>
<td>126</td>
<td>0.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>28</td>
<td>0.87</td>
<td>0.392</td>
<td>0.2945</td>
<td>170</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Split 2</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>86</td>
<td>0.90</td>
<td>0.90</td>
<td>0.364</td>
<td>0.279</td>
<td>777</td>
<td>0.90</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Calibration</td>
<td>28</td>
<td>0.78</td>
<td>0.75</td>
<td>0.448</td>
<td>0.348</td>
<td>96</td>
<td>0.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>28</td>
<td>0.89</td>
<td>0.393</td>
<td>0.312</td>
<td>205</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Split 3</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Training</td>
<td>86</td>
<td>0.92</td>
<td>0.92</td>
<td>0.293</td>
<td>0.219</td>
<td>810</td>
<td>0.92</td>
<td>2</td>
<td>6</td>
</tr>
<tr>
<td>Calibration</td>
<td>28</td>
<td>0.89</td>
<td>0.87</td>
<td>0.385</td>
<td>0.321</td>
<td>207</td>
<td>0.87</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Validation</td>
<td>28</td>
<td>0.86</td>
<td>0.501</td>
<td>0.3459</td>
<td>155</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data were randomly split three times into training (60%), calibration (20%) and validation sets (21%). Table 1 The training and calibration sets were used for building the QSAR model and the validation set was used as external set for estimating the predictive potential of the developed model. In order to develop a model with a good predictive potential, the preferable parameters of the Monte Carlo optimization, the threshold ($T^*$) and the number of epochs ($N^*$) that give the best statistics for the calibration set should be defined.

The threshold is a tool for classifying codes as both rare (and thus likely less reliable features, probably introducing noise into the model) and common features, which are used by the model and labelled as active. The optimal descriptors are calculated with the correlation weights (CW) only of active features, excluding those related to rare ones.

The Nepoch is the number of cycles (sequence of modifications of correlation weights for all codes involved in model development) for the optimization. For modelling activity of potential p-ERK inhibitors the descriptor of correlation weights (DCW) was calculated as follows: DCW (Threshold, Nepoch) = \[\sum CW(Sk) + \sum CW(SSk) + \sum CW(BOND) + \sum CW(HALO) + \sum CW(NOSP)\] (1)

where CW(SAk) is correlation weight for a molecular or structural feature (SAk) extracted from k-th SMILES; The endpoint pIC\textsubscript{50} is a function of DCW according to the following equation: \[\text{Endpoint} = C_0 + C_1 \times DCW (T, \text{Nepoch})\] (2)

Where $C_0$ and $C_1$ are the intercept and the slope for the training and calibration set, respectively.

The role of molecular features or structural attributes (SAk) extracted by SMILES can be defined by the value of CWs: SAk with positive CW are promoters of pIC\textsubscript{50}; SAk with negative CWs are responsible for the endpoint decrease; if there are both positive and negative values of CW (Sk), then that SAk has an undefined role. The applicability domain of the models can be defined according to the distribution of the molecular attributes extracted from the k-th SMILES.

A SMILES falls into the applicability domain if the following condition occurs:

\[\text{Defect SMILES } < 2 * (\text{Defect SMILES})^-\] (3)

Where Defect SMILES is the measure of the statistical (probabilistic) quality of molecular features extracted from k-th SMILES, and (Defect SMILES )^-\ is the average of these values for the training and calibration sets. Defect SMILES is calculated as follows:

\[\text{Defect SMILES } = \sum \text{Defect(\{SA\}_k)}\] (4)

\[\text{Defect(\{SA\}_k)} = |P_{\text{TRN(SAk)}} - P_{\text{TST(SAk)}}|/|N_{\text{TRN(SAk)}} + N_{\text{TST(SAk)}}|\] (5)

Where $P_{\text{TRN(SAk)}}$ is the probability of the presence of SAk in the SMILES of the training set and $P_{\text{TST(SAk)}}$ is the probability of SAk in the SMILES of the calibration set. $N_{\text{TRN(SAk)}}$ is the number (frequency) of SMILES containing SAk in the training set and $N_{\text{TST(SAk)}}$ is the number of SMILES containing SAk in the calibration set. The ideal situation occurs when the probabilities of SAk are the same in the training and in the calibration sets (Defect (SAk) is equal to zero). Conversely, if SAk is absent in the calibration set, the Defect (SAk) is maximal. Thus, the results from equation (4) can be used to classify the active attributes, and the Defect SMILES defines the domain of applicability for SMILES.

### Results and discussion

The models for three random splits are the following:

Split 01: \[pIC_{50} = 1.4942 (± 0.0237) + 0.0379 (± 0.0001) \times DCW (3, 3)\] (07)

Split 02: \[pIC_{50} = -1.0822 (± 0.0222) + 0.0342 (± 0.0001) \times DCW (5, 4)\] (08)

Split 03: \[pIC_{50} = -1.8219 (± 0.0236) + 0.0510 (± 0.0002) \times DCW (2, 6)\] (09)

Figure 3(a), Figure 3(b), Figure 3(c) shown a graphs plotting of the calculated versus experimental pIC\textsubscript{50} values.
The values of the parameters of statistical analysis of p-ERK inhibition models built on the three random distributions of data into training, calibration, and validation sets are collected in Table 2. n is the number of compounds in each set; \( r^2 \) is the coefficient of determination; \( q^2 \) is cross validated \( r^2 \); MAE is mean absolute error; s is the root-mean-square error; F is the Fischer F ratio; T and N are the preferable values for Threshold and Number of epochs, respectively. CR\(_p^2\) indicates if the models developed are obtained by chance or not, based on the Y-randomization test. For an acceptable model, CR\(_p^2\) should be greater than 0.5\(^{[27,28]}\). \( r^2 \) range from 0.78 to 0.92, and \( q^2 \) range from 0.75 to 0.92 (for training, calibration and test sets). According to the criteria defined by Tropsha\(^{[26-29]}\), a model has high predictive power if the following conditions are fulfilled:

\[
\begin{align*}
q^2 &> 0.5 \\
r^2 &> 0.6 \\
(r^2 - r^0)/r^2 &< 0.1 \\
(r^2 - r^0)/r^2 &< 0.1 \\
0.85 \leq k \leq 1.15 & \text{ or } 0.85 \leq k^\prime \leq 1.15 \\
|r^0 - r^0'| &< 0.3
\end{align*}
\]

Where \( r^0 \) and \( r^0' \) are squared correlation coefficients for regression through the origin, calculated between predicted versus experimental values and between experimental versus predicted values, k and k’ are the slopes in the former and later cases respectively; \( q^2 \) is calculated for the training sets, while all other criteria are calculated for the validation sets\(^{[26]}\). Table 3 contains the values for the criteria iii, iv, and v. Y-randomization test for all models showed that these are not chance correlations, since the CR\(_p^2\) is larger than 0.5 (Table 2). All the models successfully fulfill the criteria proposed by Tropsha\(^{[26-29]}\), a model has high predictive power if the following conditions are fulfilled:

\[
\begin{align*}
q^2 &> 0.5 \\
r^2 &> 0.6 \\
(r^2 - r^0)/r^2 &< 0.1 \\
(r^2 - r^0)/r^2 &< 0.1 \\
0.85 \leq k \leq 1.15 & \text{ or } 0.85 \leq k^\prime \leq 1.15 \\
|r^0 - r^0'| &< 0.3
\end{align*}
\]

Table 2: The numerical values of \((r^2 - r^0)/r^2; (r^2 - r^0)/r^2; k; k^\prime; \text{ and } |r^0 - r^0'|\).

<table>
<thead>
<tr>
<th>Split</th>
<th>( r^2 )</th>
<th>( r^2 )</th>
<th>( k )</th>
<th>( k^\prime )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.0244</td>
<td>0.0000</td>
<td>0.0000</td>
<td>0.9909</td>
</tr>
<tr>
<td>2</td>
<td>0.0320</td>
<td>0.0021</td>
<td>0.0018</td>
<td>0.9966</td>
</tr>
<tr>
<td>3</td>
<td>0.0136</td>
<td>0.0017</td>
<td>0.0014</td>
<td>0.9878</td>
</tr>
</tbody>
</table>

The list of all SAk, with the correlation weights for the three splits from the Monte Carlo optimization process of the built QSAR model is given in Table S2, S3 and S4. Based on the best model in validation obtained from split 2, some SMILES-based descriptors have positive influence on \( pIC_{50} \) and therefore cause an increase of \( pIC_{50} \) value\(^{[32-35]}\): =........... (Molecule containing double bond), N........... (Molecule containing nitrogen atom), O........... (Molecule containing oxygen atom), c...O....... (Molecule contains oxygen atom bonded to aromatic carbon atom), c...N....... (Molecule containing nitrogen atom bonded to aromatic carbon atom), c...S....... (Molecule containing sulfur atom bonded to aromatic carbon atom), C........... (Molecule containing sulfur atom bonded to aromatic carbon atom), BOND10000000: (molecule contains fluorine atom) etc.

The mechanistic interpretation obtained from the analysis of SAk can be used in the search and computer aided design of novel p-ERK inhibitors candidates with desired \( pIC_{50} \) values. Figure 4 shows the structures of proposed kinase inhibitors derivatives, the structure Sorafenib was used as a template for molecular design by added the SMILES notation c...S....... , BOND10000000 and c...O....... defined as promoters of increase. The molecule E has a higher value of \( pIC_{50} \) in comparison to other molecules.

Figure 4: The molecular design of possible kinase inhibitors using the QSAR model of split 2 and SAks calculated using the Monte Carlo method.
In this work three QSAR models were developed to predict V600EBRAF-dependent ERK phosphorylation of 142 molecules as kinase inhibitors, using SMILES based optimal descriptors. The models showed acceptable predictive capability on three random split of data. The analysis of the structural features (obtained from SMILES) with their positive and negative correlation weights allowed designing possible pERK inhibitors with an increased activity compared to Sorafenib.

The activity modulating effect of the structural features may serve to draw up preliminary hypothesis of novel anticancer drug candidates, and also to provide a better understanding the drugs mechanisms of action.

Acknowledgements
This research has been partially done during a research visit of the Corresponding author to Politecnico di Milano and Mario Negri Institute in Italy, in the PROSIL Life + research project of the European Union (LIFE12 ENV/IT/000154).

Conflict of Interest
The authors have no conflict of interest to declare.

Supplementary material
Table S1. Experimental and calculated of pIC50 p-ERK inhibitors derivatives designed by using the results of QSAR model obtained in this study.

Table S2. Correlation weights for DCW calculation for split 1; obtained in three probes of the Monte Carlo optimization.

Table S3. Correlation weights for DCW calculation for split 2; obtained in three probes of the Monte Carlo optimization.

Table S4. Correlation weights for DCW calculation for split 3; obtained in three probes of the Monte Carlo optimization.

Table 4: SMILES notation and pIC50 values calculated by using model of split 2 for kinase inhibitors derivatives designed by using the results of QSAR model obtained in this study.

<table>
<thead>
<tr>
<th>Molecules</th>
<th>SMILES notation</th>
<th>pIC50 (Expr.)</th>
<th>pIC50 (calc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>CNC(=O)C1=CC(OC2=CC=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>/</td>
<td>4.85</td>
</tr>
<tr>
<td>Sorafenib</td>
<td>CNC(=O)C1=CC(OC2=CC=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>5.585</td>
<td>5.3407</td>
</tr>
<tr>
<td>B</td>
<td>CNC(=O)C1=CC(OC2=CC=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>/</td>
<td>5.6258</td>
</tr>
<tr>
<td>C</td>
<td>CNC(=O)C1=CC(OC2=CC(OC)=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>5.9108</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>CNC(=O)C1=CC(OC2=CC(SC)=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>/</td>
<td>5.9694</td>
</tr>
<tr>
<td>Regorafenib</td>
<td>CNC(=O)C1=CC(OC2=CC=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>5.9965</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>CNC(=O)C1=CC(OC2=C(OC)C=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>/</td>
<td>6.5764</td>
</tr>
</tbody>
</table>

Supplementary materials section contains technical details of the models for splits examined in this work.

Conclusion
In this work three QSAR models were developed to predict V600EBRAF-dependent ERK phosphorylation of 142 molecules as kinase inhibitors, using SMILES based optimal descriptors. The models showed acceptable predictive capability on three random split of data. The analysis of the structural features (obtained from SMILES) with their positive and negative correlation weights allowed designing possible pERK inhibitors with an increased activity compared to Sorafenib.

The activity modulating effect of the structural features may serve to draw up preliminary hypothesis of novel anticancer drug candidates, and also to provide a better understanding the drugs mechanisms of action.

Acknowledgements
This research has been partially done during a research visit of the Corresponding author to Politecnico di Milano and Mario Negri Institute in Italy, in the PROSIL Life + research project of the European Union (LIFE12 ENV/IT/000154).

Conflict of Interest
The authors have no conflict of interest to declare.

Supplementary material
Table S1. Experimental and calculated of pIC50 p-ERK inhibitors derivatives designed by using the results of QSAR model obtained in this study.

Table S2. Correlation weights for DCW calculation for split 1; obtained in three probes of the Monte Carlo optimization.

Table S3. Correlation weights for DCW calculation for split 2; obtained in three probes of the Monte Carlo optimization.

Table S4. Correlation weights for DCW calculation for split 3; obtained in three probes of the Monte Carlo optimization.

Table 4: SMILES notation and pIC50 values calculated by using model of split 2 for kinase inhibitors derivatives designed by using the results of QSAR model obtained in this study.

<table>
<thead>
<tr>
<th>Molecules</th>
<th>SMILES notation</th>
<th>pIC50 (Expr.)</th>
<th>pIC50 (calc.)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>CNC(=O)C1=CC(OC2=CC=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>/</td>
<td>4.85</td>
</tr>
<tr>
<td>Sorafenib</td>
<td>CNC(=O)C1=CC(OC2=CC=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>5.585</td>
<td>5.3407</td>
</tr>
<tr>
<td>B</td>
<td>CNC(=O)C1=CC(OC2=CC=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>/</td>
<td>5.6258</td>
</tr>
<tr>
<td>C</td>
<td>CNC(=O)C1=CC(OC2=CC(OC)=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>5.9108</td>
<td></td>
</tr>
<tr>
<td>D</td>
<td>CNC(=O)C1=CC(OC2=CC(SC)=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>/</td>
<td>5.9694</td>
</tr>
<tr>
<td>Regorafenib</td>
<td>CNC(=O)C1=CC(OC2=CC=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>5.9965</td>
<td></td>
</tr>
<tr>
<td>E</td>
<td>CNC(=O)C1=CC(OC2=C(OC)C=C(NC(=O)NC3=CC(=C(CI)C=C3)C=F)F)C=C2)=CC=N1</td>
<td>/</td>
<td>6.5764</td>
</tr>
</tbody>
</table>


