Chapter 15

QSPR/QSAR Analyses by Means of the CORAL Software: Results, Challenges, Perspectives

Andrey A. Toropov
IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Italy

Alla P. Toropova
IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Italy

Emilio Benfenati
IRCCS-Istituto di Ricerche Farmacologiche Mario Negri, Italy

Orazio Nicolotti
Università degli Studi di Bari “Aldo Moro”, Italy

Angelo Carotti
Università degli Studi di Bari “Aldo Moro”, Italy

Karel Nesmerak
Charles University in Prague, Czech Republic

Aleksandar M. Veselinović
University of Niš, Serbia

Jovana B. Veselinović
University of Niš, Serbia

Pablo R. Duchowicz
Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Argentina

Daniel Bacelo
Universidad de Belgrano, Argentina

Eduardo A. Castro
Instituto de Investigaciones Fisicoquímicas Teóricas y Aplicadas INIFTA (UNLP, CCT La Plata-CONICET), Argentina

Bakhtiyor F. Rasulev
Jackson State University, USA

Danuta Leszczynska
Jackson State University, USA

Jerzy Leszczynski
Jackson State University, USA

DOI: 10.4018/978-1-4666-8136-1.ch015
ABSTRACT

In this chapter, the methodology of building up quantitative structure—property/activity relationships (QSPRs/QSARs)—by means of the CORAL software is described. The Monte Carlo method is the basis of this approach. Simplified Molecular Input-Line Entry System (SMILES) is used as the representation of the molecular structure. The conversion of SMILES into the molecular graph is available for QSPR/QSAR analysis using the CORAL software. The model for an endpoint is a mathematical function of the correlation weights for various features of the molecular structure. Hybrid models that are based on features extracted from both SMILES and a graph also can be built up by the CORAL software. The conceptually new ideas collected and revealed through the CORAL software are: (1) any QSPR/QSAR model is a random event; and (2) optimal descriptor can be a translator of eclectic information into an endpoint prediction.

INTRODUCTION

Almost seventy years ago one of the areas of theoretical chemistry named the quantitative structure – property/activity relationships (QSPRs/QSARs) was established (Wiener, 1947, pp. 17-20; Wiener, 1948, pp. 425-430). However, if one considers the publication of Hammett’s Free-energy relationship (Hammett, 1935, pp. 125–136) as the first exercise in the QSPR/QSAR analyses, then the introduction of this area of theoretical chemistry should be moved back by eighty years. This is obviously an evidence of the long lasting importance of the QSAR approaches. Such significance is due to their wide applications. The main target of QSPR/QSAR analyses is the prediction of the numerical data related to various endpoints (physicochemical in the case of QSPR; and biochemical, ecological, medicinal, etc., in the case of QSAR).

One can consider three periods of evolution in this field.

1. The first period can be defined as the construction of molecular descriptors and establishing of their correlation with various endpoints (Bonchev & Trinajstic, 1977; Balaban, 1982, pp. 399-404; Bonchev, 1995, pp. 137-156; Diudea & Gutman, 1998; Ivanciuc, Ivanciuc, & Balaban, 1998; Randic, deAlba, & Harris, 1998; Castro, Tueros, & Toropov, 2000; Randic & Basak, 2001; Nikolic, Kovacevic, Milicevic, & Trinajstic, 2003; Ren, 2003, pp. 29-39; Gonzalez, Teran, Tejeira, & Gonzalez-Moa, 2005; Melagrazi, Afantitis, Sarimveis, Iglesissi-Markopouloua, & Supuran, 2006).

2. The second period can be defined as the construction of molecular descriptors, establishing of their correlation with an endpoint for the compounds of “visible” training set with the further testing of the predictive potential of this correlation with “invisible” compounds (i.e. compounds which are not involved in building up of the model) involved in the test set (Cronin et al., 2002; Golbraikh & Tropsha, 2002; Taskinen & Yliiruusi, 2003; Hemmateenejad, 2004, pp. 475–485; Duchowicz, Castro, Fernandez, & Gonzalez, 2005; Oberg, 2005, pp. 2189–2200; Afantitis et al., 2006; Coi et al., 2006; Leonard & Roy, 2006; Pan, Jiang, & Wang, 2007; Roy & Roy, 2008; Porto, Souza, Junkes, Yunes, & Heinzen, 2008; Toropova, Toropov, Benfenati, Leszcynska, & Leszczynski, 2010; Toropov et al., 2012a; Toropov, Toropova, Raska Jr, Benfenati, & Gini, 2012b; Nesmerak, Toropov, Toropova, & Kohoutova, 2013; Papa, van derWal, Arnot, & Gramatica, 2014).

3. The third period can be defined as building up of the QSPR/QSAR models according to OECD principles (Kruhlak, Contrera, Daniel Benz, & Matthews, 2007; Kar & Roy, 2010, 2012; Putz, Ionascu, Putz, & Ostafe, 2011; de Melo, 2012, pp. 213–222; Sahigara et al., 2012; Toropova, 2004).
According to the OECD principles, the QSPR/QSAR model should be characterized by:

1. A defined endpoint;
2. An unambiguous algorithm;
3. A defined domain of applicability;
4. Clear measures of predictive potential; and
5. A mechanistic interpretation, if possible (OECD, 2007).

It was during the second period of the methodology development, when optimal descriptors calculated with the molecular graph were suggested as a tool to solve the QSPR/QSAR tasks (Randic & Basak, 1999, 2001; Randic & Pompe, 2001; Toropov & Toropova, 2002; da Silva Junkes, Arruda, Yunes, Porto, & Heinzen, 2005). In fact, majority of molecular descriptors are not expected to correlate with a certain endpoint, but they can be involved in building up model for arbitrary endpoint. However, the optimal descriptor is aimed to correlate with a certain endpoint via a one-variable correlation.

Interestingly, simplified molecular input-line entry system (SMILES) is an attractive alternative for representation of the molecular structure by graph (Weininger, 1988, pp. 31-36; Weininger, 1990, pp. 237-243; Weininger, Weininger, & Weininger, 1989). There are databases available on the Internet where the molecular structure is represented by SMILES (Toropov, Toropova, Raska Jr, Leszczynska, & Leszczynski, 2014). Under such circumstances, the development of QSPR/QSAR approaches, where SMILES is the representation of the molecular structure becomes an attractive direction of research work in the field of the QSPR/QSAR theory and applications. The CORAL software is a tool of the QSPR/QSAR analyses with representation of the molecular structure by SMILES (Toropova, Toropov, Benfenati, Leszczynska, & Leszczynski, 2010d; Toropova et al., 2011a,b). It is to be noted that the CORAL software is able to extract from the SMILES various graph theoretical invariants such as vertex degree and extended connectivity of higher orders (Toropov & Toropova, 2002), as well as invariants for the graph of atomic orbitals (Toropov & Toropova, 2001).

The description of:

1. The organization of the CORAL software,
2. Results obtained using this software,
3. Advantages and disadvantages of the approach, as well as
4. Discussion of possible modifications of this software are the main aims of the chapter.

**Background**

The basic paradigm of the QSPR/QSAR prediction can be expressed as the following:

\[ \text{Endpoint} = F(\text{Molecular Structure}) \]

The classic mathematical representation of the molecular structure can be done by means of the adjacency matrix \( A(G) \) (Gutman, 1988, pp. 93-94; Gutman, 1997, pp. 281-287). The hydrogen-suppressed graph (HSG) for 3-pentanol (CAS 584-02-1) can be represented as seen in Figure 1.
The hydrogen-suppressed graph is characterized by the following adjacency matrix $A_{(HSG)}$:

$$
\begin{array}{cccccccc}
 1 & 2 & 3 & 4 & 5 & 6 & \delta_k & \\
1 & 0 & 1 & 0 & 0 & 0 & 1 & \\
2 & 1 & 0 & 1 & 0 & 0 & 2 & \\
3 & 0 & 1 & 0 & 1 & 0 & 1 & \\
4 & 0 & 0 & 1 & 0 & 1 & 2 & \\
5 & 0 & 0 & 0 & 1 & 0 & 1 & \\
6 & 0 & 0 & 1 & 0 & 0 & 1 & \\
\end{array}
$$

Majority of the molecular descriptors are a mathematical function of the $A_{(HSG)}$, in particular the Randic index (Randic, 1975) is calculated as

$$
X = \sum_{(i,j)\text{edge}} (\delta_i \delta_j)^{-0.5}.
$$

The pioneer version of the optimal descriptor (Randic & Basak, 1999; Randic, 2001) is represented by the calculation of classic descriptors using the modified adjacency matrix $A^{*}_{(HSG)}$:

$$
X^* = \sum_{(i,j)\text{edge}} (\delta_i^* \delta_j^*)^{-0.5}.
$$

The next step in the evolution of the optimal descriptors is the application instead of numerical contributions of vertex degrees ($\delta_j$) the so-called correlation weight of the presence of a certain vertex degree, together with the correlation weight of a certain chemical element ($A_k$):

$$
X_{CW} = \sum_k CW(A_k) + \sum_k CW(\delta_k).
$$
The correlation weights are calculated by the Monte Carlo method. The numerical data on the correlation weights being used for calculation of the $^0X_{cw}$ should provide maximum of correlation coefficient between the descriptor and a certain endpoint for the training set of compounds. Having the numerical data one can calculate the $^0X_{cw}$ for compounds of the training set and define the model:

$$
\text{Endpoint} = C_0 + C_1 \times ^0X_{cw}
$$

(4)

However, the predictive potential of the model calculated with Equation 4 should be checked up with external test set.

The representation of the molecular structure in databases by adjacency matrix is not convenient and non economical, in comparison to the representation by SMILES string. Consequently, one can define the descriptor similar to the above-mentioned $^0X_{cw}$ which can be calculated with SMILES:

$$
DCW = F(\text{SMILES})
$$

(5)

The simplest way is to define the descriptor as a mathematical function of SMILES atoms, i.e. as a mathematical function of each character of SMILES, except situations where two symbols cannot be examined separately (e.g. ‘Cl’, ‘Br’, etc.). In the case of 3-pentanol with SMILES= “CCC(O)CC”, the scheme can be illustrated as the following:

$$
DCW = CW('C') + CW('C') + CW('C') + CW('O') + CW('O') + CW('C') + CW('C').
$$

(6)

The scheme expressed by Equation 6, can be useful for relatively simple physicochemical properties, such as the normal boiling point of a certain class of organic compounds (alkanes, alkylbenzenes, ethers, etc.). However, in the case of mixture of various classes of organic compounds, or in the case of more complex endpoints (toxicity, carcinogenicity, etc.), the scheme gives poor statistics for both the “visible” training set and for the “invisible” test set. Fortunately, the approach can be improved if in addition to correlation weights of SMILES atoms, combinations of two and / or three SMILES-atoms would be involved in the model.

However, in the cases of two or three SMILES atoms some paradoxical situations occur. For example, SMILES fragment “CN” and “NC” (similarly, “CCN” and “NCC”) will be examined as different molecular features. Fortunately, one can avoid these cases if useses normalized combines of SMILES atoms. A simple way is to select order which obey principle: the first and last characters are placed according to their ASCII codes (Toropov, Toropova, & Benfenati, 2010b). In addition, SMILES attributes which are fragments of molecular structure, the global invariants of the molecular structure are involved to build up a model. These global attributes are a mathematical function of presence (absence) of various molecular features which are represented in Table 1.
**Table 1. The scheme of extraction of SMILES atoms and other SMILES attributes in order to build up a model**

<table>
<thead>
<tr>
<th>SMILES Attribute</th>
<th>Example of the Representation for the CORAL Software</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_k$</td>
<td>SMILES-atoms, i.e. one symbol or two symbols which cannot be examined separately, e.g. 'C' and 'Cl': this information is represented by sequences of twelve symbols:</td>
</tr>
<tr>
<td></td>
<td><img src="image1" alt="Example Image" /></td>
</tr>
<tr>
<td>$SS_k$</td>
<td>A combination of two SMILES-atoms ‘CC’ and ‘CN’: this information is represented by sequences of twelve symbols:</td>
</tr>
<tr>
<td></td>
<td><img src="image2" alt="Example Image" /></td>
</tr>
<tr>
<td>$SSS_k$</td>
<td>A combination of three SMILES-atoms ‘CNC’ and ‘C#N’; this information is represented by sequences of twelve symbols:</td>
</tr>
<tr>
<td></td>
<td><img src="image3" alt="Example Image" /></td>
</tr>
<tr>
<td><strong>BOND</strong></td>
<td>The presence / absence of double (‘=’), triple (‘#’), and stereo chemical (‘@’) bonds, e.g. if SMILES = “CCC(O)CC”</td>
</tr>
<tr>
<td></td>
<td><img src="image4" alt="Example Image" /></td>
</tr>
<tr>
<td><strong>NOSP</strong></td>
<td>Presence (absence) of nitrogen, oxygen, sulphur, and phosphorus, e.g. if SMILES=“CCC(O)CC”</td>
</tr>
<tr>
<td></td>
<td><img src="image5" alt="Example Image" /></td>
</tr>
<tr>
<td><strong>HALO</strong></td>
<td>Presence (absence) of fluorine, chlorine, bromine, and iodine atoms, e.g. if SMILES=“ClCC(=O)CCI”</td>
</tr>
<tr>
<td></td>
<td><img src="image6" alt="Example Image" /></td>
</tr>
<tr>
<td><strong>PAIR</strong></td>
<td>Simultaneous presence of two SMILES-atoms from the list: F, Cl, Br, I, N, O, S, P, #, =, and @; e.g. if SMILES=“ClCC(=O)CCI” the following pairs will be extracted:</td>
</tr>
<tr>
<td></td>
<td><img src="image7" alt="Example Image" /></td>
</tr>
</tbody>
</table>
The above-mentioned Monte Carlo optimization in the case of the unlimited number of epochs will provide maximum of correlation coefficient for the training set. However (Figure 2), a maximum of correlation coefficient between optimal descriptor and endpoint for a certain number of the epochs exists for the external test set. This amount of epochs is preferable in calculations, because when the number of epochs reaches such a value the obtained model is characterized by a good predictive potential.

Vice versa, the increase of threshold is accompanied by decrease of the correlation coefficient for the training set, but again, the threshold which provide maximum correlation coefficient for the test set exists. This threshold is preferable from practical point of view. Consequently, in order to prepare a good model by means of the CORAL software, one should define preferable values of the threshold and the number of epochs of the Monte Carlo optimization (T* and N*) (Toropova et al., 2011a).

Table 2. The representations of the molecular structure which can be used to build up a model by means of the CORAL software

<table>
<thead>
<tr>
<th>The Representation of the Molecular Structure</th>
<th>Graphical Interpretation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SMILES CN(C)OC</td>
<td><img src="image1.png" alt="Graph" /></td>
</tr>
<tr>
<td>Hydrogen-Suppressed Graph, HSG</td>
<td><img src="image2.png" alt="Graph" /></td>
</tr>
<tr>
<td>Hydrogen-Filled Graph, HFG</td>
<td><img src="image3.png" alt="Graph" /></td>
</tr>
<tr>
<td>Graph of Atomic Orbitals, GAO</td>
<td><img src="image4.png" alt="Graph" /></td>
</tr>
</tbody>
</table>
Table 2 contains the list of possible representations of the molecular structure which can be used to build up a model. Thus, the CORAL software is able to provide the following types of QSPR/QSAR models:

1. **SMILES Based Models:**

   \[ SMILES_{DCW}(\text{Threshold, } N_{\text{epoch}}) = \alpha \sum CW(S_k) + \beta \sum CW(SS_k) + \gamma \sum CW(SSS_k) + x \cdot CW(\text{NOSP}) + y \cdot CW(\text{HALO}) + z \cdot CW(\text{BOND}) + t \cdot CW(\text{PAIR}) \] ; \hspace{1cm} (7)

2. **Models Based on the Representation by Molecular Graphs:** These are HSG, or HFG, or GAO (Table 2).

   \[ Graph_{DCW}(\text{Threshold, } N_{\text{epoch}}) = \sum CW(A_k) + \alpha \sum CW(0EC_k) + \beta \sum CW(1EC_k) + \gamma \sum CW(2EC_k) + \delta \sum CW(3EC_k) \] ; \hspace{1cm} (8)

   In Equations 7 and 8, coefficients \( \alpha, \beta, \gamma, \delta, t, x,y, \) and \( z \) can be 1 (yes) or 0 (no). Various combines of their values give possibility to define certain version of the optimal descriptor;

3. **The Hybrid Models which Involves Molecular Feature Extracted from Both Smiles and Graph:**

   \[ Hybrid_{DCW}(\text{Threshold, } N_{\text{epoch}}) = SMILES_{DCW}(\text{Threshold, } N_{\text{epoch}}) + Graph_{DCW}(\text{Threshold, } N_{\text{epoch}}) \] ; \hspace{1cm} (9)

4. Last but not least: the CORAL software is able to provide models for phenomena where eclectic information is the only basis for prediction, e.g. property and activity of proteins (Toropov, Toropova, Raska Jr, Benfenati, & Gini, 2012b) as well as endpoint related to nanomaterials (Toropova & Toropov, 2013). The generalized form of the optimal descriptor which can be used for the above-mentioned situations is the following:

   \[ DCW(\text{Threshold, } N_{\text{epoch}}) = \Sigma CW(CI_k) \] \hspace{1cm} (10)

   where \( CI_k \) represent codes of available eclectic data (e.g. concentration, size, condition of synthesis, technological details, etc.).

**Selected Technical Details Related to the CORAL Software**

There are some limitations of the SMILES in CORAL software. The maximal length for SMILES is 500 symbols; the maximal number of atoms (or atomic orbitals in the case of the graph of atomic orbitals) is equal to 300; the maximal number of compounds (which are involved in building up model) is equal to 50,000.
Development of a Model by Means of the CORAL Software Is Based on the Following Principles

1. Molecular structure of each compound can be represented by molecular features which are extracted from SMILES and/or from molecular graph;
2. There are local and global molecular features which can be extracted from SMILES. The local features represent some fragments (or individual atoms). The global features are some indexes which characterized molecules as whole.
3. Building up of QSPR/QSAR model for an arbitrary split into the training and validation sets can be examined as a random event.
4. The statistical quality of each QSPR/QSAR model is a mathematical function that represents split of the data into the “visible” training set and “invisible” test set.
5. The average statistical quality of QSPR/QSAR models that is obtained for several splits into training and test sets is more robust criterion for the estimation of an approach than the statistical quality for solely one split.
6. The average statistical quality of a model for external “invisible” validation sets is more significant information related to the quality of the model than the average statistical quality for “visible” (i.e. substances involved in building up model) sub-training and calibration sets.

Every model that includes geometry-dependent molecular descriptors usually involves a relatively expensive calculation of the optimum molecular geometry, involving high computational costs and long CPU time. In this context, the conformation-independent 0D, 1D and 2D-QSPR methods emerge as an alternative approach for developing models based on constitutional and topological molecular features of compounds (Duchowich, Comelli, Ortiz, & Castro, 2012; Talevi et al., 2012). The exclusion of 3D-structural aspects also avoids problems associated with ambiguities, resulting from an incorrect geometry optimization output due to the existence of compounds in various conformational states. Such kind of problems may also lead to the vanishing of predictive capability of the QSPR/QSAR when applied for the prediction of an external test set of compounds. In the realms of the CORAL approach, the calculated flexible descriptor is a molecular descriptor which depends both on the molecular structure and the property under analysis, but does not explicitly depend on details from the 3D-molecular geometry. In previous QSPR/QSAR studies the importance of this methodology has been shown. Such approach is able to provide models having a comparable or sometimes better quality than the ones found by searching the best descriptors in a pool containing thousands of 0D-3D descriptors (Mullen, Duchowicz, & Castro, 2011; Garcia et al., 2011; Ibezim, Duchowicz, Ortiz, & Castro, 2012).

THE COLLECTION OF QSPR/QSAR MODELS

Over the years the CORAL software has been used for QSPR/QSAR analyses of a group of various endpoints. The results are listed in the following sections.
**Water Solubility of Organic Compounds**

Data on water solubility of 1311 substances, i.e. their CAS number, SMILES, and the values of negative logarithm of water solubility – log S (mol / L) were taken from the web site of Virtual Computational Chemistry Laboratory (http://www.vcclab.org/lab/alogps/). These substances were distributed by means of five random splits into the sub-training set (~55%), calibration set (~25%), and test set (~20%). The statistical quality of the CORAL models (Table 3) is better than the statistical quality of the models suggested earlier in the literature for the same large database (Huuskonen, 2000, pp. 773-777; Tetko, Tanchuk, Kasheva, & Villa, 2001; Yan & Gasteiger, 2003). The best prediction is characterized by \( n_{\text{test}}=21, r_{\text{test}}^2 =0.91, s_{\text{test}}=0.63 \) (Huuskonen, 2000, pp. 773-777).

**Solubility of [C60] and [C70] Fullerene Derivatives**

The numerical data on the solubility of [C60] and [C70] fullerene derivatives in chlorobenzene in mg/mL were taken from the literature (Troshin et al., 2009). Unfortunately, since at present, the QSPR analysis of these data by other methods is not available we are not able to compare the obtained results with various approaches.

**Octanol/Water Partition Coefficient of Antineoplastic Drugs**

The numerical data on octanol/water partition coefficient (logP) for 55 organic compounds have been taken from the literature (Wishart at al., 2008). The CORAL model (Table 3) is better than the model described in the literature (Fouchécourt, Béliveau, & Krishnan, 2001) for the same data (n=55, \( r^2 =0.8180 \)).

**Half-Wave Potential**

The half-wave oxidation/reduction potential is very useful characteristic of electronic properties of the chemical compounds. The first SMILES-based QSPR analysis of half-wave potentials was performed for 40 benzoazines (Toropov, Nesmerak, Raska Jr., Waiss, & Palat, 2006). The statistical quality this approach (\( r^2 = 0.882 \)) is fully comparable with classical approach based on Hammett constants (\( r^2 = 0.897 \); Nesmerak, Nemec, Sticha, Waiss, & Palat, 2005). Other example of SMILES-based QSPR of half-wave potential using CORAL software is analysis of experimental data for 23 derivatives of \( N\)-benzylsalicylthioamides. From the statistical point of view, the model is comparable with data obtained in the experiment (Nesmerak, Toropov, Toropova, Kohouťová, & Waiss, 2013).

**Rate Constants of Reactions between Organic Aromatic Pollutants and Hydroxyl Radical**

The degradability of 78 compounds expressed in terms of second-order degradation rate constants, \( K_{\text{OH}} \) between OH- and each targeted aromatic compound have been taken from the literature (Kušić, Rasulev, Leszczyńska, Leszczyński, & Koprivanac, 2009). Decimal logarithm of the \( K_{\text{OH}} \) was evaluated as the endpoint for the QSPR models. The model from the above-mentioned work (Kušić, Rasulev, Leszczyńska, Leszczyński, & Koprivanac, 2009) is characterized by \( r_{\text{test}}^2 =0.760 \). Hence the statistics of the CORAL model (Table 3) are better, because for four random splits into the training and test sets for the CORAL models (Toropov et al., 2012a) the average \( r_{\text{test}}^2 =0.8253 \).
<table>
<thead>
<tr>
<th>Endpoint</th>
<th>( N_{\text{train}} )</th>
<th>( r^2_{\text{train}} )</th>
<th>( s_{\text{train}} )</th>
<th>( N_{\text{test}} )</th>
<th>( r^2_{\text{test}} )</th>
<th>( s_{\text{test}} )</th>
<th>( R_m^2 )</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Physicochemical Parameters</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Water solubility of organic compounds</td>
<td>721</td>
<td>0.9231</td>
<td>0.565</td>
<td>261</td>
<td>0.9381</td>
<td>0.511</td>
<td>0.8963</td>
<td>Toropov et al., 2013a</td>
</tr>
<tr>
<td>Solubility of [C60] and [C70] fullerene derivatives</td>
<td>18</td>
<td>0.758</td>
<td>17.6</td>
<td>9</td>
<td>0.925</td>
<td>12.5</td>
<td>0.902</td>
<td>Toropova et al., 2011b</td>
</tr>
<tr>
<td>Octanol/water partition coefficient of antineoplastic drugs</td>
<td>22</td>
<td>0.9306</td>
<td>0.601</td>
<td>12</td>
<td>0.9776</td>
<td>0.727</td>
<td>-</td>
<td>Toropov et al., 2010a</td>
</tr>
<tr>
<td>Half-wave potential</td>
<td>9</td>
<td>0.7333</td>
<td>0.021</td>
<td>5</td>
<td>0.7295</td>
<td>0.028</td>
<td>0.8294</td>
<td>Toropov et al., 2011a</td>
</tr>
<tr>
<td>Rate Constants of Reactions Between Organic Aromatic Pollutants and Hydroxyl Radical</td>
<td>38</td>
<td>0.8207</td>
<td>0.183</td>
<td>20</td>
<td>0.8033</td>
<td>0.145</td>
<td>0.6731</td>
<td>Toropov et al., 2012a</td>
</tr>
<tr>
<td><strong>Biochemical Activities</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Toxicity towards <em>Daphnia magna</em></td>
<td>149</td>
<td>0.7006</td>
<td>1.04</td>
<td>89</td>
<td>0.7680</td>
<td>0.878</td>
<td>0.7413</td>
<td>Toropov et al., 2012a</td>
</tr>
<tr>
<td>Toxicity in rats</td>
<td>344</td>
<td>0.7662</td>
<td>0.468</td>
<td>126</td>
<td>0.7237</td>
<td>0.315</td>
<td>0.6388</td>
<td>Toropov et al., 2011a</td>
</tr>
<tr>
<td>Toxicity of binary mixtures</td>
<td>14</td>
<td>0.9584</td>
<td>0.167</td>
<td>10</td>
<td>0.9362</td>
<td>0.200</td>
<td>0.7164</td>
<td>CORAL: Models of toxicity of binary mixtures. Chemometrics and Intelligent Laboratory Systems, 119</td>
</tr>
<tr>
<td>Carcinogenicity</td>
<td>170</td>
<td>0.6278</td>
<td>0.871</td>
<td>61</td>
<td>0.7576</td>
<td>0.602</td>
<td>-</td>
<td>Toropova et al., 2010b</td>
</tr>
<tr>
<td>Carcinogenicity of drugs</td>
<td>1088</td>
<td>0.8700</td>
<td>1.46</td>
<td>376</td>
<td>0.8520</td>
<td>1.53</td>
<td>0.835</td>
<td>Toropova &amp; Toropov, 2014</td>
</tr>
<tr>
<td>Anticancer agents Murine P388 Leukemia</td>
<td>75</td>
<td>0.7688</td>
<td>0.48</td>
<td>25</td>
<td>0.8025</td>
<td>0.49</td>
<td>-</td>
<td>Toropov et al., 2010c</td>
</tr>
<tr>
<td>Anticancer agents SK-MEL-5 cell line UACC-62 cell line</td>
<td>19</td>
<td>0.9506</td>
<td>0.14</td>
<td>6</td>
<td>0.752</td>
<td>0.36</td>
<td>-</td>
<td>Mullen et al., 2011</td>
</tr>
<tr>
<td>Anti-malaria agents aryl-piperazine derivatives</td>
<td>16</td>
<td>0.9960</td>
<td>0.040</td>
<td>6</td>
<td>0.8836</td>
<td>-</td>
<td>-</td>
<td>Ibezim et al., 2012</td>
</tr>
<tr>
<td>Anti-malaria agents 4(1-H)-quinolone derivatives</td>
<td>20</td>
<td>0.7009</td>
<td>0.575</td>
<td>10</td>
<td>0.6130</td>
<td>0.89</td>
<td>0.55</td>
<td>Toropov et al., 2013b</td>
</tr>
<tr>
<td>Anti-HIV-1</td>
<td>7</td>
<td>0.8413</td>
<td>0.309</td>
<td>5</td>
<td>0.8379</td>
<td>0.642</td>
<td>-</td>
<td>Toropov et al., 2010d</td>
</tr>
<tr>
<td>High-affinity 5-HT1A receptors</td>
<td>40</td>
<td>0.9457</td>
<td>0.223</td>
<td>18</td>
<td>0.9473</td>
<td>0.298</td>
<td>0.8325</td>
<td>Veselinovic et al., 2013a</td>
</tr>
<tr>
<td>Calcium Channel-Antagonistic Effect</td>
<td>50</td>
<td>0.8507</td>
<td>0.706</td>
<td>22</td>
<td>0.9269</td>
<td>0.521</td>
<td>0.8558</td>
<td>Veselinovic et al., 2013b</td>
</tr>
<tr>
<td>Anticonvulsant agents</td>
<td>46</td>
<td>0.7627</td>
<td>0.192</td>
<td>5</td>
<td>0.7001</td>
<td>0.313</td>
<td>0.6428</td>
<td>Garro Martinez et al., 2011</td>
</tr>
<tr>
<td>Alkaloid toxicity</td>
<td>26</td>
<td>0.83</td>
<td>0.19</td>
<td>8</td>
<td>0.70</td>
<td>0.44</td>
<td>-</td>
<td>Turabekova et al., 2014</td>
</tr>
<tr>
<td>Membrane damage TiO2 nanoparticles</td>
<td>10</td>
<td>0.9893</td>
<td>0.025</td>
<td>5</td>
<td>0.9647</td>
<td>0.066</td>
<td>0.90</td>
<td>Toropova &amp; Toropov, 2013</td>
</tr>
</tbody>
</table>
Chromatographic Retention Data

The retention factors of 41 derivatives of 1-phenyl-5-benzylsulfanyltetrazole in reverse-phase HPLC were measured for various compositions of mobile-phase. The retention characteristics of the linear solvent strength model were calculated using the linear relationship between logarithm of retention factor and volume fraction of organic modifier in mobile phase. The logarithms of the solute retention factor extrapolated to a mobile phase composition with 0% organic modifier, and constants for a given solute in a given chromatographic system (i.e. the slopes of the linear regressions) were used as endpoints. The one-variable QSPRs developed show that the CORAL software is a tool for building up a robust quantitative structure-retention relationships (QSRR) model for retention characteristics (Nesmerak, Toropov, & Toropova, 2014).

Toxicity towards *Daphnia magna*

The descriptions of toxicity of organic chemicals related to 48 h *Daphnia magna* toxicity expressed in negative decimal logarithm of the dose that kills 50% of organisms i.e. pLC50 were taken from the literature (Kar & Roy, 2010). The data set covers range of octanol/water partition coefficient from −2 to 8. According the values, it could be more possibly logarithms of octanol/water partition coefficient. The range of toxicity (daphnia) is from 0.46 to 10.09. In regard to the chemical domain, the data set includes hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives; iso-thiocyanates; thiols; phosphorothionate and phosphate esters; and halogenated derivatives. The statistical quality of the model described in the original article is the following: n=222, $r^2=0.695$ (training set), and n=75, $r^2_{pred}=0.741$, $R_m^2=0.707$ (test set). Thus, the statistical quality of the Coral model (Table 3) is at least comparable with the model suggested in the above-mentioned work (Kar & Roy, 2010).

Toxicity in Rats

Rat toxicity data (LD50, in mg/kg, oral exposure, n=689) was taken from the U.S. Library of Medicine (http://toxnet.nlm.nih.gov/). The log10[1/LD50] has been used as endpoint. Unfortunately, information on QSAR analysis of these data by other approaches is not available. However, comparison of optimal descriptors and Multiple Linear Regression (MLR) has shown (Toropov, Rasulev, & Leszczynski, 2008) that the models for rat toxicity based on optimal descriptors calculated with SMILES have better statistical characteristics than one-, two-, three-, and four-variable MLR models based on the topological and quantum mechanical descriptors.

Toxicity of Binary Mixtures

The numerical data on the toxicity of binary mixtures was taken from the literature (Zhang, Zhou, Yang, & Wang, 2007). The toxicity is expressed as pEC50 (i.e. negative decimal logarithm log[1/EC50]), logarithm of the inverse of the effective concentration required to bring about a 50% decrease in light emission, for *Photobacterium phosphoreum* (T3 mutation). The statistical quality of four-variables model (calculated with involvement of the quantum mechanical descriptors) suggested in the literature (Zhang, Zhou, Yang, & Wang, 2007) for the toxicity of the same 50 binary mixtures is the following: n=50, $r^2=0.85$, s=0.270. Consequently, the CORAL model for this endpoint provides better results (Table 3).
Carcinogenicity

Experimental values for carcinogenicity were taken from the web site of United States Environmental Protection Agency (http://www.epa.gov/nct/dsstox/sdf_cpdbas.html). Carcinogenicity is expressed as the potency dose that induces cancer in rats (TD50, in mg/kg body weight). These values have been converted into mmol/kg body weight. The log(TD50) was selected as endpoint for modeling. The correlation coefficient of the QSAR for carcinogenicity of 45 benzene derivatives amounts to about $r^2 = 0.7$ (Vracko, 1997, pp.1037–1043). Another QSAR model for rodent carcinogenicity nitroso compounds obtained by multiple linear regression analysis (Helguera, Cordeiro, Pérez, Combes, & González, 2008) has the following statistical characteristics: $n=48$, $r^2 = 0.859$, $s=0.361$, $F=42$ (training set) and $n=6$, $Q^2 = 0.71$, $s=0.488$ (test set). The CORAL model for carcinogenicity (Table 3) should be consider as comparable with the above-mentioned models, since the number of examined compounds for the CORAL model is considerable larger.

Carcinogenicity of Drugs

Numerical data on carcinogenic potentials of the selected 1464 organic compounds (chemical domain which includes hydrocarbons, aliphatic alcohols, phenols, ethers, and esters; anilines, amines, nitriles, nitroaromatics, amides, and carbamates; urea and thiourea derivatives, isothiocyanates, thiols, phosphate esters, and halogenated derivatives) are expressed by DF (Discriminate Function). The range of DF is from -9.91 to 9.86. Positive value of DF is an indicator of carcinogenic compound, negative value of DF is an indicator of non-carcinogenic compound. The predictive potentials of QSAR models for the same series of compounds suggested in the literature are characterized by $n_{test} = 732$, $r^2_{test} = 0.713$ (Kar & Roy, 2011); and $n_{test} = 732$, $r^2_{test} = 0.77$ (Duchowicz, Comelli, Ortiz, & Castro, 2012). Thus the statistical quality of the CORAL model is better than those available in the literature (Table 3).

Anticancer Agents (Murine P388 Leukemia)

The decimal logarithm log(1/IC50), where IC50 represents the concentration of the agent necessary to reduce cell viability by 50% against Murine P388 Leukemia (in vitro cytotoxic activity) is considered as the endpoint. Numerical data on this endpoint was taken from the literature (Atanasova, Ilieva, & Galabov, 2007). The QSAR model suggested in the published work (Atanasova et al., 2007) is characterized by $n_{test} = 25$; $r^2_{test} = 0.700$. Thus, predictive potential of the CORAL model for the anticancer activity is better that the previously published results.

Anticancer Agents (SK-MEL-5 Cell Line and UACC-62 Cell Line)

The observed cell anti-proliferative activities of the triphenylmethyl containing compounds inducing death in the human melanoma cell lines SK-MEL-5 and UACC-62 are extracted from the literature (Palchaudhuri, Nesterenko, & Hergenrother, 2008; Palchaudhuri & Hergenrother, 2008). The models calculated with MLR approach (Mullen, Duchowicz, & Castro, 2011) are characterized by $n_{train} = 17$, $r^2_{train} = 0.781$, $s_{train} = 0.16$, $n_{test} = 6$, $r^2_{test} = 0.243$, $s_{test} = 0.44$ (SK-MEL-5); and $n_{train} = 19$, $r^2_{train} = 0.859$, $s_{train} = 0.17$, $n_{test} = 4$, $r^2_{test} = 0.722$, $s_{test} = 0.30$ (UACC-62). It can be compared to the results obtained using CORAL approach, included in the Table 3.
Anti-Malaria Agents (Aryl-Piperazine Derivatives)

The experimental inhibitory concentrations (IC50) in micromolar units of aryl-piperazine derivatives against the chloroquine resistant strains W2 and FCR3 and against the chloroquine-sensitive strains D10 and NF54 are extracted from the literature (Molyneaux, Krugliak, Ginsburg, & Chibale, 2005). The QSAR calculated with optimal descriptors gave possibility to suggest some perspective molecular structures of anti-malaria agents.

Anti-Malaria Agents (4(1-H)-Quinolone Derivatives)

A series of 53 endochin analogs (4(1-H)-quinolone derivatives) with anti-malarial activity against the clinically relevant multidrug resistant malarial strain TM-90-C2B has been taken from the literature (Ojha & Roy, 2011). The statistical characteristics of the model suggested in the published work (Ojha & Roy, 2011) are the following: \( n_{\text{train}} = 39, r^2_{\text{train}} = 0.797, s_{\text{train}} = 0.517; n_{\text{test}} = 14, r^2_{\text{test}} = 0.808 \). Thus, the statistical quality of the CORAL model is comparable with the above-mentioned model (Ojha & Roy, 2011).

Anti-HIV-1

The binding affinity values of fullerene derivatives pEC50 were taken from the literature (Durdagi, Mavromoustakos, & Papadopoulos, 2008). The QSAR analysis of these data has shown that even for the case of small set of data, the Monte Carlo technique gives reproducible results. In spite of relative lower statistical quality of the CORAL model for the endpoint, this model provides satisfactory prediction for a several splits of the data into the training and test sets. Thus, the optimal descriptors should be considered as reasonable alternative for the 3D QSAR analysis.

High-Affinity 5-HT1A Receptors

A dataset of 88 arylpiperazines considered as high affinity 5-HT1A receptor ligands, to which the in vitro affinity values (as measured by inhibition constants, Ki) were collected from the literature (Martinez-Esparza et al., 2001a,b) was analyzed. The CORAL software is able to be an efficient tool to build up a robust model for affinity of 5-HT1A receptor ligands of arylpiperazines. The suggested modeling process for affinity of 5-HT1A receptor ligands is based on the representation of the molecular structure by SMILES and on the application of experimental data. The predictive potential of the applied approach was tested with four random splits of the data into the training and test sets. In addition, the SMILES attributes, which are promoters of increase/decrease ligand binding affinity to 5-HT1A receptor, were identified. The mechanistic interpretation of these models from probabilistic point of view was suggested.

Calcium Channel-Antagonistic Effect

A dataset of 72 1,4-dihydropyridine derivatives with values for pIC50 was taken from the literature (Gecen, Saripinar, Yannaz, & Şahin, 2012). Almost identical statistical quality of the QSAR models has been obtained for four different random splits of data. The mechanistic interpretation of these models in terms of promoter increase / decrease of the endpoint was suggested.
QSPR/QSAR Analyses by Means of the CORAL Software

Anticonvulsant Agents

The experimental information on the antiepileptic activities ED50 of the molecular structures was obtained from the literature (Eddington et al., 2002, 2003; Edafiogho, Ananthalakshmi, & Kombian, 2006; Cox, Gao, Raje, Scott, & Eddington, 2001a; Cox, Scott, Gao, Raje, & Eddington, 2001b; Vamecq, Lambert, Poupaert, Masereel, & Stables, 1998). The MLR analysis gives preferable accuracy of the two ED50 models (Garro Martinez, Duchowicz, Estrada, Zamarbide, & Castro, 2011) $r^2_{\text{test}}=0.856$ and $r^2_{\text{test}}=0.948$. For this case the CORAL model provides worse results (Table 3).

Anesthetic Activity of Alkaloids

The duration of anesthesia (related to protein binding of a drug) and the onset time (determined by the $pK_a$) are important characteristics in assessment of local anesthetic agents. In the study (Turabekova et al., 2014) we utilized toxicity, duration, and onset of action as endpoints to construct QSAR models for the series of 34 diterpenoid alkaloids characterized by local anesthetic activity using GA-MLRA/PLS and SMILES-based optimal descriptors approach (CORAL). SMILES-based optimal descriptor approach application resulted in models of relatively improved both statistical fit of toxicity and duration of anesthesia models ($r^2\approx85\%$ and $r^2\approx95\%$, respectively) and predicting ability ($\approx72\%$) for all three endpoints. This approach appeared to be more sensitive to structural peculiarities of molecules than regression methods that have to be taken into account in potential data splitting.

Membrane Damage TiO$_2$ Nanoparticles

Large databases related to physicochemical and biological activity of nanomaterials remain scarce. Under such circumstances, the possible way to build up a predictive model is an analysis of eclectic available data on various phenomena related to nanomaterials. As a result of exchange of classic paradigm $\text{Endpoint} = F(\text{molecular structure})$ by paradigm $\text{Endpoint} = F(\text{eclectic information})$, an analogy of classic QSPR/QSAR approach applied to nanomaterials can be obtained (Toropova & Toropov, 2013). The data obtained using this approach are provided in the Table 3.

SOLUTIONS AND RECOMMENDATIONS

The variety of QSPR/QSAR models developed with the CORAL software indicates that this approach can be an efficient tool of the QSPR/QSAR analyses. However, it is to be noted, there are problems related to utilization of the CORAL software.

Problem 1: The involvement of combinations of two and three SMILES atoms in building up a model can lead to appearance of SMILES attributes (Table 1) with unclear physical meaning, e.g.:
Problem 2: The approach is able to detect three-atom fragments as basis for the mechanistic interpretation. In reality, an active fragment can involve larger number of atoms.

Problem 3: If all SMILES contain certain, an indeed active fragment, this fragment cannot be recognized by the CORAL software as an active one.

Problem 4: If an active fragment is rare, the fragment cannot be recognized by CORAL software as an active one.

However, a preliminary analysis and classification (filtration) of dataset can help to solve these problems, at least partially.

FUTURE RESEARCH DIRECTIONS

The recently published works have shown, that the CORAL software can be a tool of QSPR/QSAR analyses for other endpoints in addition to results listed in Table 3 (Singh and Gupta, 2014; Achary, 2014, pp. 73-90; Worachartcheewan, Nantasenamat, Isarankura-Na-Ayudhya, & Prachayasittikul, 2014). QSAR Prediction Reporting Format (QPRF) (version 1.1, May 2008) available on the Internet http://ihcp.jrc.ec.europa.eu/our_labs/predictive_toxicology/qsar_tools/qrf/QPRF_version_1%201_DEREK_SS.pdf is multi-component system. Most likely, the system can be useful and attractive for the case of standardized situations. However, in order to be widely applied as a research tool, it is preferable to use OECD principles together with the following probabilistic rules:

1. A QSPR/QSAR model should be represented for a group of splits of data into the “visible” training set (compounds involved in the modeling process) and the “invisible” test set (compounds which are not involved in the modeling process);
2. The splits should be random and different (they should be far from identity). The Monte Carlo technique in general, and the CORAL software in particular, can be basis for the QSAR Prediction Reporting Format, at least for the first steps of the QSPR/QSAR analyses of ecologic, pharmacologic and other systems of endpoints which are characterized by high complexity and wide prevalence.
QSPR/QSAR Analyses by Means of the CORAL Software

CONCLUSION

The CORAL software can be considered as an universal tool for the QSPR/QSAR analysis. As all software packages the CORAL has certain advantages and disadvantages. The software provides the following options:

1. QSPR/QSAR analysis for data represented in format described at http://www.insilico.eu/coral;
2. The analyses should be carried out using a number of various data splits;
3. The software can be involved in build up models for phenomena characterized by eclectic information. Thus, the most important advantages of the approach are:
   a. The possibility of analysis of a QSPR/QSAR as a random event (not unique action); and
   b. The possibility of the translation of eclectic information into endpoint prediction.

ACKNOWLEDGMENT

We thank the EC project NANOPUZZLES (Project Reference: 309837). D.L. and J.L. acknowledge support from the National Science Foundation (NSF/CREST HRD-0833178). E.A.C., D.E.B. and P.R.D. thank for the financial support provided by the National Research Council of Argentina (CONICET) PIP11220100100151 project and to Ministerio de Ciencia, Tecnología e Innovación “Productiva for the electronic library facilities.”

REFERENCES


QSPR/QSAR Analyses by Means of the CORAL Software


**KEY TERMS AND DEFINITIONS**

**Correlation Weight:** A coefficient which should be added to optimal descriptor if a given SMILES attribute or graph invariant takes place in the given SMILES or in the given molecular graph.

**Global SMILES Attribute:** A characteristic of molecules in whole, e.g. the presence of nitrogen together with oxygen, presence of double bonds and cycles, the presence of nitrogen and sulphur and absence of oxygen, and others (Table 1).

**Graph of Atomic Orbitals (GAO):** Molecular graph where vertexes are the representation of atomic orbitals (e.g. 1s2, 2p3, 3d7, etc.).

**Hydrogen-Filled Graph (HFG):** Molecular graph where presence of hydrogen atoms is taken into account.

**Hydrogen-Suppressed Graph (HSG):** Molecular graph where presence of hydrogen atoms is ignored.

**Local SMILES Attribute:** Fragment of SMILES line which contains one (‘C’, ‘N’, ‘=’, etc.) or two symbols (‘Cl’, ‘Br’, ‘@@’, etc.) which cannot be examined separately.

**Monte Carlo Optimization:** The calculation of the numerical data on the correlation weights which give maximal value of the correlation coefficient between an endpoint and the optimal descriptor.

**Optimal Descriptor:** A descriptor which is calculated for a substances with simplified molecular input-line entry system (SMILES) or with molecular graph using correlation weights for SMILES attributes or for graph invariant (vertex degree, valence shells, etc.).