Chem Biol Drug Des 2011; **77:** 471–476

Research Article

doi: 10.1111/j.1747-0285.2011.01117.x

CORAL Software: QSAR for Anticancer Agents

Emilio Benfenati, Andrey A. Toropov*, Alla P. Toropova, Alberto Manganaro and Rodolfo Gonella Diaza

Istituto di Ricerche Farmacologiche Mario Negri, Via La Masa 19, 20156 Milano, Italy
**Corresponding author: Andrey A. Toropov, aatoropov@yahoo.com*

CORrelations And Logic (CORAL at http://www.insilico.eu/coral) is freeware aimed at establishing a quantitative structure – property/activity relationships (QSPR/QSAR). Simplified molecular input line entry system (SMILES) is used to represent the molecular structure. In fact, symbols in SMILES nomenclatures are indicators of the presence of defined molecular fragments. By means of the calculation with Monte Carlo optimization of the so called correlation weights (contributions) for the above-mentioned molecular fragments, one can define optimal SMILES-based descriptors, which are correlated with an endpoint for the training set. The predictability of these descriptors for an external validation set can be estimated. A collection of SMILES-based models of anticancer activity of 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridines for different splits into training and validation set which are calculated with the CORAL are examined and discussed. Good performance has been obtained for three splits: the r^2 ranged between 0.778 and 0.829 for the sub-training set, between 0.828 and 0.933 for the calibration set, and between 0.807 and 0.931 for the validation set.

Key words: anticancer activity, optimal descriptor, QSAR, SMILES

Received 28 May 2010, revised 16 February 2011 and accepted for publication 6 March 2011

There are a number of systems for the establishing of the quantitative structure – property /activity relationships (QSPR/QSAR) based on different collections of molecular descriptors (1–4). The use of large databases, especially databases that are available via the Internet, is typical in modern natural sciences. The majority of the Internet databases oriented on molecular properties are based on the representation of the molecular structure by simplified molecular input line entry system (SMILES) (5–8). Thus, the development of molecular descriptors that are calculated directly from SMILES is an attractive scenario of the QSPR/QSAR researches. The CORrelations And Logic (CORAL) software is an attempt to develop the standardized SMILES-based optimal descriptors. The aim of the present publication is the demonstration of the ability of the CORAL freeware to be a tool for the QSAR modelling. The numerical data on the anticancer activity for 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridines (9) is used to demonstrate this freeware in practice.

Most popular 'classic' approach of QSAR modelling can be formulated as the following: (i) definition of a model with compounds of the training set; and (ii) checking of the model with compounds of an external validation set. One can formulate a few questions related to the optimization of this approach. For instance, how the statistical quality of the model will be modified in case of another split into the training and validation sets? How to avoid the overtraining (i.e., how avoid the situation when a good model for the training set becomes a poor model for external substances)? How one can estimate the probability of obtaining a satisfactory and reliable model?

Algorithms that are used in the CORAL can give some solutions for the above-mentioned problems from a probabilistic point of view. In fact, CORAL is a producer of random models, which are calculated by the Monte Carlo method. A random model can be a reasonable predictor for an endpoint, if the statistical quality of this model (for both the training and validation set) can be reproduced in a sequence of attempts to build this model. Obeying to this logic, we have examined three different splits in a cascade of attempts to build the models for the anticancer activity.

In addition to the above-mentioned classic scheme, one can use the balance of correlations that is available in the CORAL. The basic idea of the balance of correlations is the split of the training set into sub-training and calibration set. The preliminary check of the model is the function of the calibration set. This preliminary check helps to avoid the overtraining. As further step to improve the predictability, the balance of correlations with ideal slopes has been examined. Slopes of the cluster at the plot of experimental versus calculated values of the endpoint on the sub-training and calibration set are ideal if their values are as equivalent as possible.

Thus, the discussion of the CORAL as a tool for QSPR/QSAR analyses is the aim of the present work, considering the specific case study of anticancer activity.

Method

Data

The concentration of the agent to reduce cell viability by 50%, against Murine P388 Leukemia IC_{50} (9) is the biological activity

Table 1: Example of preparation *S*, *SS*, and *SSS* attributes. Vacant positions are indicated by 'x' Selected SMILES is Cl.O=-C(O)C2=CN(c1nc(c(F)cc1C2=O)N3CCC(N)C3)c4ccccc4

S Zone-1 Zone-2 Zone-3	*SS* Zone-1 Zone-2 Zone-3	*SSS* Zone-1 Zone-2 Zone-3
Clxxxxxxxxxx		
.xxxxxxxxxxx	Clxx.xxxxxxx	
Oxxxxxxxxxxx	Oxxx.xxxxxxx	Oxxx.xxxClxx
=xxxxxxxxxxx	Oxxx=xxxxxxx	=xxxOxxx.xxx
Cxxxxxxxxxxx	Cxxx=xxxxxxx	Oxxx=xxxCxxx
(xxxxxxxxxxx	Cxxx(xxxxxxx	=xxxCxxx(xxx
Oxxxxxxxxxxx	Oxxx(xxxxxxx	Oxxx(xxxCxxx
(xxxxxxxxxxx	Oxxx(xxxxxxx	(xxxOxxx(xxx
Cxxxxxxxxxxx	Cxxx(xxxxxxx	Oxxx(xxxCxxx
2xxxxxxxxxxx	Cxxx2xxxxxxx	2xxxCxxx(xxx
=xxxxxxxxxxx	=xxx2xxxxxxx	Cxxx2xxx=xxx
Cxxxxxxxxxxx	Cxxx=xxxxxxx	Cxxx=xxx2xxx
Nxxxxxxxxxxx	NxxxCxxxxxxx	NxxxCxxx=xxx
(xxxxxxxxxxx	Nxxx(xxxxxxx	CxxxNxxx(xxx
cxxxxxxxxxxx	cxxx(xxxxxxx	cxxx(xxxNxxx
1xxxxxxxxxxx	cxxx1xxxxxxx	1xxxcxxx(xxx
nxxxxxxxxxxx	nxxx1xxxxxxx	nxxx1xxxcxxx
cxxxxxxxxxxx	nxxxcxxxxxxx	cxxxnxxx1xxx
(xxxxxxxxxxx	cxxx(xxxxxxx	nxxxcxxx(xxx
cxxxxxxxxxxx	cxxx(xxxxxxx	cxxx(xxxcxxx
(xxxxxxxxxxx	cxxx(xxxxxxx	(xxxcxxx(xxx
Fxxxxxxxxxxx	Fxxx(xxxxxxx	cxxx(xxxFxxx
(xxxxxxxxxxx	Fxxx(xxxxxxx	(xxxFxxx(xxx
cxxxxxxxxxxx	cxxx(xxxxxxx	cxxx(xxxFxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxx(xxx
1xxxxxxxxxxx	cxxx1xxxxxxx	cxxxcxxx1xxx
Cxxxxxxxxxxx	Cxxx1xxxxxxx	cxxx1xxxCxxx
2xxxxxxxxxxx	Cxxx2xxxxxxx	2xxxCxxx1xxx
=xxxxxxxxxxx	=xxx2xxxxxxx	Cxxx2xxx=xxx
Oxxxxxxxxxxx	Oxxx=xxxxxxx	Oxxx=xxx2xxx
(xxxxxxxxxxx	Oxxx(xxxxxxx	=xxxOxxx(xxx
Nxxxxxxxxxxx	Nxxx(xxxxxxx	Oxxx(xxxNxxx
3xxxxxxxxxxx	Nxxx3xxxxxxx	3xxxNxxx(xxx
Cxxxxxxxxxxx	Cxxx3xxxxxxx	Nxxx3xxxCxxx
Cxxxxxxxxxxx	CxxxCxxxxxxx	CxxxCxxx3xxx
Cxxxxxxxxxxx	CxxxCxxxxxxx	CxxxCxxxCxxx
(xxxxxxxxxxx	Cxxx(xxxxxxx	CxxxCxxx(xxx
Nxxxxxxxxxxx	Nxxx(xxxxxxx	Nxxx(xxxCxxx
(xxxxxxxxxxx	Nxxx(xxxxxxx	(xxxNxxx(xxx
Cxxxxxxxxxxx	Cxxx(xxxxxxx	Nxxx(xxxCxxx
3xxxxxxxxxxx	Cxxx3xxxxxxx	3xxxCxxx(xxx
(xxxxxxxxxxx	3xxx(xxxxxxx	Cxxx3xxx(xxx
cxxxxxxxxxxx	cxxx(xxxxxxx	cxxx(xxx3xxx
4xxxxxxxxxxx	cxxx4xxxxxxx	4xxxcxxx(xxx
cxxxxxxxxxxx	cxxx4xxxxxxx	cxxx4xxxcxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxx4xxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxxcxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxxcxxx
cxxxxxxxxxxx	cxxxcxxxxxxx	cxxxcxxxcxxx
4xxxxxxxxxxx	cxxx4xxxxxxx	cxxxcxxx4xxx

examined in this work. The values of decimal logarithm log(1/IC50) or pEC50 are the endpoint that is modelled by the SMILES-based optimal descriptors. The SMILES nomenclatures used in this work have been generated by ACD/ChemSketch.[a]

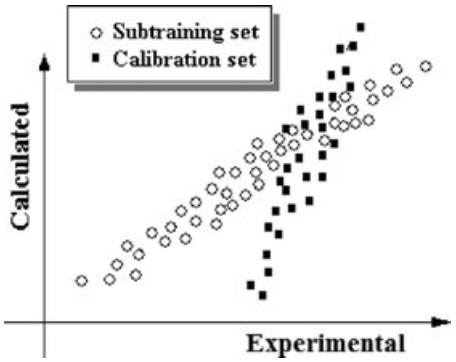

Figure 1: Good correlations that are accompanied by different slopes for the sub-training set and the calibration set in plots of experimental versus calculated values of an endpoint. Classic scheme balance of correlations balance of correlations with ideal slopes.

Descriptors

Optimal SMILES-based descriptors (10) are calculated as the following:

$$\mathrm{DCW(Threshold)} = \alpha \sum_{k=1}^{E} W(S_k) + \beta \sum_{k=1}^{E-1} W(SS_k) + \gamma \sum_{k=1}^{E-2} W(SSS_k) \quad (1)$$

where S_k, SS_k, and SSS_k are one-, two-, and three-element SMILES attributes; $W(S_k)$, $W(SS_k)$, and $W(SSS_k)$ are the correlation weights of the attributes. The SMILES element is one or two symbols that should be examined as united ones (e.g., 'Cl', 'Br'). The threshold is a value used for classification of attributes into two classes: rare and active. For instance, if threshold is 5, then attributes that take place in four (or less) SMILES of the training set should be classified as rare. The correlation weights of rare attributes are blocked: their values are fixed equal to zero. The E is the number of S_k. If a SMILES is a sequence of element 'ABCDE', then the construction of S_k, SS_k, and SSS_k may be represented as the following:

$$\begin{aligned} &\mathrm{ABCDE} \rightarrow \mathrm{A, B, C, D, E}(S_k) \\ &\mathrm{ABCDE} \rightarrow \mathrm{AB, BC, CD, DE}(SS_k) \\ &\mathrm{ABCDE} \rightarrow \mathrm{ABC, BCD, CDE}(SSS_k) \end{aligned} \quad (2)$$

These are some real example: if SMILES = cccCl then S_k = (c, c, c, Cl); SS_k = (cc, cc, cCl); SSS_k = (ccc, ccCl).

The constants α, β, and γ can be used to modify the DCW(Threshold): they can be defined as either 0 or 1. The simplest version of the descriptor takes place if $\alpha = 1$, $\beta = 0$, and $\gamma = 0$. The most complex version of the descriptor takes place if $\alpha = 1$, $\beta = 1$, and $\gamma = 1$. The overtraining is a main reason for serious criticism of QSAR (11–13). In this perspective, it should be noted that the above-mentioned version of the DCW(Threshold), i.e., $\alpha = 1$, $\beta = 0$, and $\gamma = 0$, is the most robust. In other words, in this case the probability of the overtraining is minimal. However, in this case the statistical quality for the training/

Table 2: Statistical quality of the best QSAR models for the anticancer potential calculated with CORAL with different approaches and thresholds (T = 1–15). The n is the number of compounds in a set; R is correlation coefficient; s is root-mean-square error; F is Fischer F-ratio; N_{act} is the number of active (not blocked) attributes; N_{111} is the number of active SMILES attributes which take place in all sets; N_{110} is the number of SMILES attributes which take place in sub-training and calibration sets; N_{101} is the number of SMILES attributes which take place in sub-training and validation sets; N_{100} is the number of SMILES attributes which are absent in the calibration and validation sets; W (%) = 100 $[N_{111}/(N_{111} + N_{110} + N_{101} + N_{100})]$, i.e., percent of SMILES attributes which are present in the sub-training, calibration, and test set

			Sub-training set				Calibration set				Validation set								
Split	T	N_{act}	n	r^2	s	F	n	r^2	s	F	n	r^2	s	F	W (%)	N_{111}	N_{110}	N_{101}	N_{100}
Classic scheme																			
1	14	145	75	0.8828	0.339	550					25	0.5570	0.804	29	100	0	0	145	0
2	14	145	75	0.8599	0.391	448					25	0.6763	0.675	48	100	0	0	145	0
3	10	163	75	0.8469	0.399	404					25	0.8680	0.388	151	100	0	0	163	0
Balance of correlations																			
1	15	103	50	0.8291	0.411	233	25	0.9740	0.496	863	25	0.5912	0.727	33	100	103	0	0	0
2	15	97	50	0.8101	0.461	205	25	0.8128	0.521	100	25	0.7805	0.504	82	100	97	0	0	0
3	14	121	50	0.7535	0.470	147	25	0.8369	0.491	118	25	0.8563	0.404	137	100	121	0	0	0
Balance of correlations with ideal slopes																			
1	12	139	50	0.7738	0.473	164	25	0.9323	0.499	317	25	0.8118	0.484	99	100	139	0	0	0
2	5	170	50	0.8291	0.437	233	25	0.8291	0.471	112	25	0.8041	0.482	95	100	170	0	0	0
3	4	190	50	0.8358	0.384	244	25	0.8435	0.469	124	25	0.9349	0.372	331	94	178	2	7	3

sub-training set, as rule, is modest. Hence, more complex versions (e.g., α = 1, β = 1, and γ = 0 or α = 0, β = 0, and γ = 1) can also be useful for QSPR/QSAR modelling. Any combination of the constants (excepting α = 0, β = 0, and γ = 0) is available for CORAL calculations.

In the present study, SMILES elements containing one (e.g., 'c', 'C', '=') or two symbols (e.g., 'Cl', 'Br') have been examined. In general, SMILES elements containing larger number of symbols may be defined.

The modelling approach examined in this study includes three steps (14,15):

Step 1

Preparation of the list of SMILES attributes for every SMILES notation. Each SMILES attribute is a string of 12 symbols. This string is separated into three zones. The first four symbols is the zone-1; the second four symbols is the zone-2; and the third four symbols is the zone-3.

There are three categories of the SMILES attributes. The first category refers to attributes (S_k) containing sole SMILES element positioned in the zone-1; the second category includes attributes (SS_k) containing two SMILES elements positioned in zone-1 and zone-2; the third category includes attributes (SSS_k) containing three SMILES elements positioned in zone-1, zone-2, and zone-3. Table 1 contains an example of the preparation of a list of the attributes for a SMILES notation.

To avoid the situation when two different SMILES attributes are representing the same molecular fragments, for instance the '(N' and the 'N(', the elements for the SS_k and SSS_k are ranged according to their ASCII codes. Furthermore, the symbol ')' is replaced by '(', because these are representations of the same phenomenon (i.e., branch in molecular skeleton). The same takes place for '[' and ']'.

Step 2

Preparation of the completed list of the SMILES attributes that take place in the work set (i.e., totally in the training/sub-training, calibration, and test sets). The correlation weights of all SMILES attributes are set as equal to 1.

Step 3

The optimization of the correlation weights has been done by using the Monte Carlo method. The algorithm of the Monte Carlo optimization (15) has been used in two versions. The first is the traditional classic scheme: correlation weights, which produce as large as possible correlation coefficient between the DCW(Threshold) and endpoint on the training set, are calculated (10,14,15).

The second scheme, i.e., the balance of correlations is the following: available data were split into sub-training, calibration, and validation set. The target function (14,15) of the optimization for this scheme is calculated as

$$\mathrm{BC} = R + R' - \mathrm{ABS}(R - R')\,\mathrm{dR\text{-}weight} \qquad (3)$$

where R and R' are correlation coefficients between DCW(Threshold) and endpoint for the sub-training and calibration set, respectively; dR-weight is an empirical coefficient. As a rule, dR-weight = 0.1 is a satisfactory choice.

The optimization with the target function calculated with eqn 3 can lead to an unreliable model demonstrated in Figure 1.

To avoid this situation, one can use the modified version of the target function, calculated as the following:

$$IS = BC - abs(C0 + C0' + abs(C1 - C1'))dC\text{-weight} \quad (4)$$

where C0 and C0′ are intercepts for the sub-training and calibration set; C1 and C1′ are slopes for the sub-training and calibration set. The dC-weight is an empirical coefficient. Usually, a satisfactory range for this coefficient is about 0.01–0.005.

The user of CORAL can select (i) classic scheme, (ii) balance of correlations (i.e., the Monte Carlo optimization of target function calculated with eqn 3), and (iii) balance of correlations with ideal slopes (i.e., the Monte Carlo optimization of target function calculated with eqn 4). The present study is based on the target function calculated with eqn 4.

Thus, the quality of the models, i.e., correlation coefficients (R), root mean square error (s), Fischer F-ratio (F) for all sets (sub-training, calibration, and validation) is components of a mathematical function:

$$Q = F(\text{Split, DCW-version, dR-weight, dC-weight, } d_{start}, d_{precision}, N_{epoch}, \text{Threshold, Approach}) \quad (5)$$

where the different parameters for CORAL are indicated, and in particular Split is the selected separation into sub-training, calibration, and validation sets; DCW-version is related to the selected values of α, β, and γ; d_{start}, $d_{precision}$ are starting values of precision of the optimization procedure by the method of division by half; N_{epoch} is the number of epochs of the optimization; Approach is the selection of classic scheme or balance of correlations or balance of correlations with ideal slopes. The CORAL-method used in the present study adopts the following scenario: splits are 1, 2, and 3; DCW-version is ($\alpha = 1$, $\beta = 1$, $\gamma = 1$); dR-weight = 0.1; dC-weight = 0.025 (split 1); 0.03 (split 2 and 3); $d_{start} = 0.1$; $d_{precision} = 0.1$; $N_{epoch} = 30$; threshold = 1,2,…,15; and approaches are (i) classic scheme, (ii) balance of correlations, and (iii) balance of correlations with ideal slopes.

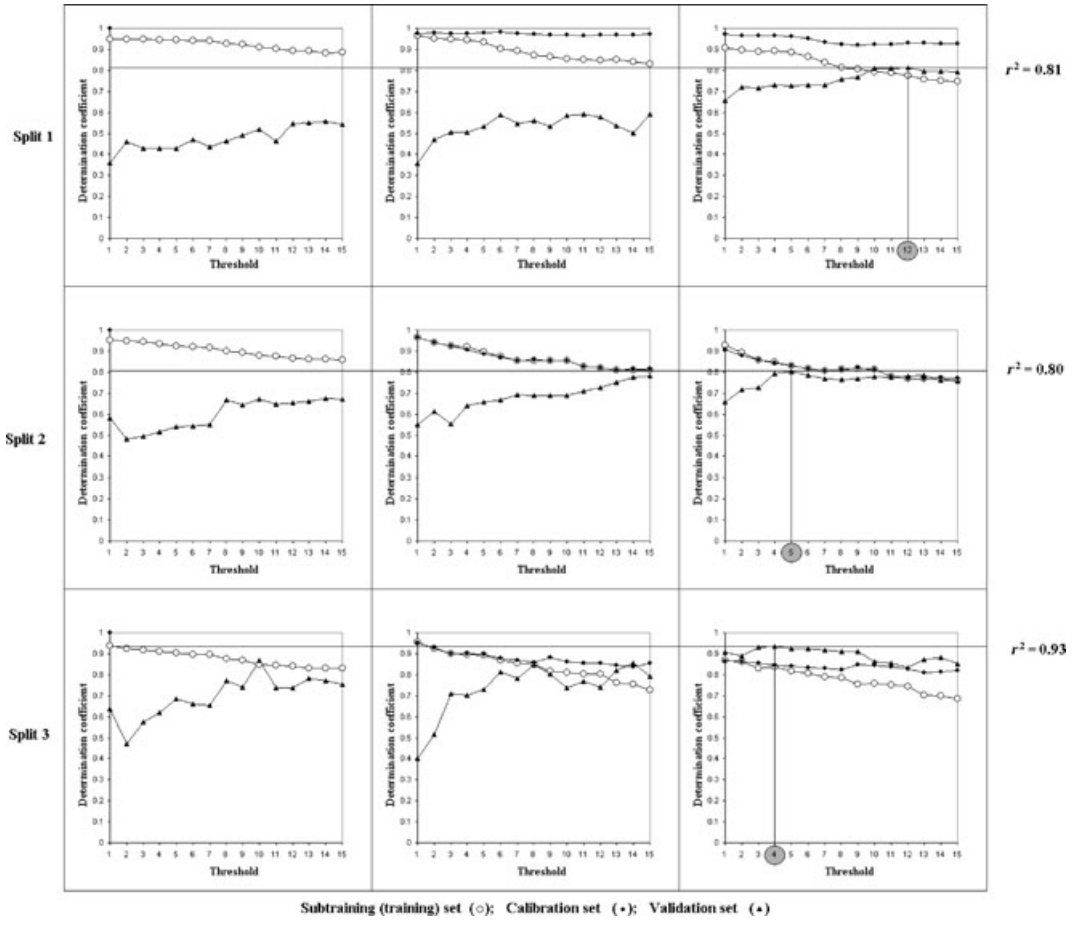

Figure 2: Statistical quality of CORAL models for the anticancer activity obtained for three splits with threshold values of 1–15. In the cases of split 1, the best predictions ($r^2 = 0.81$) takes place with threshold = 12; in the case of split 2, the best prediction takes place with threshold = 5; in the case of split 3, the best prediction takes place with threshold = 4.

Results and Discussion

Table 2 contains the parameters indicating the statistical quality of the models obtained by different versions of the methods represented by generalized eqn 5. Figure 2 graphically shows the data contained in Table 2. The dispersion of R^2 is <0.005. The dispersion of the standard error of estimation is <0.003 (in logarithm units).

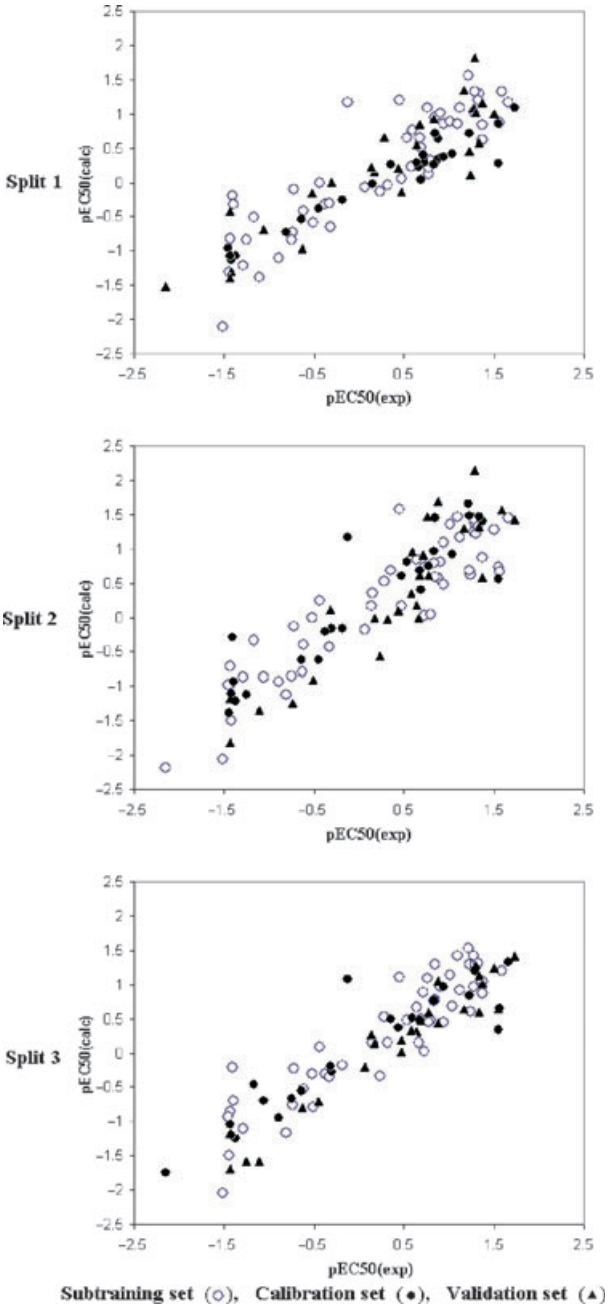

Figure 3: QSAR models for anticancer activity for three random splits 1, 2, and 3 which are calculated with eqns 6, 7 and 8, respectively.

Thus, one can see that for each split, there is a threshold that provides reasonable and reliable prediction for $\log(1/IC_{50})$. The threshold values are 12 (for split 1), 5 (for split 2), and 4 (for split 3). The balance of correlations (without of the ideal slopes, i.e., the target function calculated with eqn 3) gives satisfactory models for three splits. However, models obtained with target function calculated with eqn 4 are better (Table 2). We note that the values of threshold for models based on the target function calculated with eqns 3 and 4 are different (Table 2, Figure 2). Table S1 contains the best models for splits 1, 2, and 3. Figure 3 graphically shows these models.

Thus, CORAL has two phases of calculations: (i) definition of the optimal threshold and (ii) building up a final model with the optimal threshold. Table 2 shows results of the first-phase calculation.

The final model for split 1 (balance of correlations with ideal slopes the threshold = 12) is the following:

$$\mathrm{pEC50} = -0.1627\,(\pm 0.0105) + 0.1000\,(\pm 0.0011)\,\mathrm{DCW}(12) \quad (6)$$

n = 50, r^2 = 0.7779, q^2 = 0.7597, s = 0.469, F = 168 (sub-training set);

n = 25, r^2 = 0.9330, $R^2_{\mathrm{pred}} = 0.9241$, s = 0.493 (calibration set);

n = 25, r^2 = 0.8067, $R^2_{\mathrm{pred}} = 0.7737$, $R^2_{\mathrm{m}} = 0.8027$ (it should be >0.5); s = 0.487, (test set);

$(r^2 - r_0^2)/r^2$ = 0.0000 [it should be <0.1 (16)];

$(r^2 - r_0'^2)/r^2$ = 0.0025 [it should be <0.1 (16)];

k = 1.0984 [it should be $0.85 \le k \le 1.15$ (16)].

The statistical quality of the final model for split 2 (balance of correlations with ideal slopes the threshold = 5) are the following:

$$\mathrm{pEC50} = -0.0003\,(\pm 0.0092) + 0.0899\,(\pm 0.0008)\,\mathrm{DCW}(5) \quad (7)$$

n = 50, r^2 = 0.8287, q^2 = 0.8151, s = 0.438, F = 232 (sub-training set);

n = 25, r^2 = 0.8283, $R^2_{\mathrm{pred}} = 0.8050$, s = 0.466 (calibration set);

n = 25, r^2 = 0.8160, $R^2_{\mathrm{pred}} = 0.7797$, $R^2_{\mathrm{m}} = 0.6712$ [it should be >0.5 (17)], s = 0.468, (test set);

$(r^2 - r_{0^2})/r^2$ = 0.0386 [it should be <0.1 (16)];

$(r^2 - r_0'^2)/r^2$ = 0.0119 [it should be <0.1 (16)];

k = 1.0283 [it should be $0.85 \le k \le 1.15$ (16)].

The final model for the split 3 (balance of correlations with ideal slopes the threshold = 4) is the following:

$$\text{pEC50} = -0.0113\,(\pm 0.0094) + 0.0851\,(\pm 0.0008)\,\text{DCW(4)} \quad (8)$$

n = 50, r^2 = 0.8285, q^2 = 0.8136, s = 0.392, F = 232 (sub-training set);

n = 25, r^2 = 0.8475, $R^2_{\text{pred}} = 0.8210$, s = 0.473 (calibration set);

n = 25, r^2 = 0.9305, $R^2_{\text{pred}} = 0.9186$, $R^2_{\text{m}} = 0.6825$ [it should be >0.5 (17)]; s = 0.372, (test set);

$(r^2 - r_0^2)/r^2$ = 0.0763 [it should be <0.1 (16)];

$(r^2 - r_0'^2)/r^2$ = 0.0552 [it should be <0.1 (16)];

k = 1.0660 [it should be $0.85 \le k \le 1.15$ (16)].

Table S1 shows the models calculated with eqns 6, 7 and 8. One can reproduce these results using CORAL freeware available on the Internet.[b]

Conclusions

(i) One can estimate CORAL models as quite satisfactory for their statistical performance; (ii) balance of correlations with ideal slopes gives better prediction for anticancer activity than the balance correlations without of the ideal slopes; and (iii) the results obtained with balance of correlations with ideal slopes are reproducible with different splits.

Acknowledgment

The authors thank the European Commission for financial support (the contract OSIRIS). The authors also express their gratitude to Dr. L. Cappelini and Dr. G. Bianchi for technical assistance and to Dr J. Baggot for the English revision.

References

1. Devillers J. (1996) Genetic Algorithms in Molecular Modeling. London: Academic Press, Ltd.
2. Todeschini R., Consonni V. (2000) Handbook of Molecular Descriptors. Weinheim, New York: Wiley-VCH.
3. Devillers J., Balaban A.T., editors (2000) Topological Indices and Related Descriptors in QSAR and Drug Design. Amsterdam: Gordon&-Breach.
4. Todeschini R., Consonni V. (2003) DRAGON Software for the Calculation of Molecular Descriptors, web version 3.0 for Windows.
5. Weininger D. (1988) SMILES, a chemical language and information system 1: Introduction and encoding rules. J Chem Inf Comput Sci;28:31–36.
6. Weininger D., Weininger A., Weininger J.L. (1989) SMILES. 2: Algorithm for generation of unique SMILES notation. J Chem Inf Comput Sci;29:97–101.
7. Weininger D. (1990) SMILES. 3. Depict. Graphical depiction of chemical structures. J Chem Inf Comput Sci;30:237–243.
8. Toropov A.A., Toropova A.P., Mukhamedzhanova D.V., Gutman I. (2005) Simplified molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-property relationships (QSPR). Indian J Chem;44:1545–1552.
9. Atanasova M., Ilieva S., Galabov B. (2007) QSAR analysis of 1,4-dihydro-4-oxo-1-(2-thiazolyl)-1,8-naphthyridines with anticancer activity. Eur J Med Chem;42:1184–1192.
10. Toropov A.A., Toropova A.P., Benfenati E. (2009) QSAR modelling for mutagenic potency of heteroaromatic amines by optimal SMILES-based descriptors. Chem Biol Drug Des;73:301–312.
11. Doweyko A.M. (2004) 3D-QSAR illusions. J Comput Aided Mol Des;18:587–596.
12. Doweyko A.M. (2008) QSAR: dead or alive? J Comput Aided Mol Des;22:81–89.
13. Johnson A.R. (2008) The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). J Chem Inf Model;48:25–26.
14. Toropov A.A., Rasulev B.F., Leszczynski J. (2008) QSAR modeling of acute toxicity by balance of correlations. Bioorg Med Chem;16:5999–6008.
15. Toropov A.A., Toropova A.P., Benfenati E. (2009) Additive SMILES-based carcinogenicity models: Probabilistic principles in the search for robust predictions. Int J Mol Sci;10:3106–3127.
16. Golbraikh A., Tropsha A. (2002) Beware of q2! J Mol Graph Model;20:269–276.
17. Roy P.P., Roy K. (2008) On some aspects of variable selection for partial least squares regression models. QSAR Comb Sci;27: 302–313.

Notes

[a]ACD/ChemSketch Freeware, version 11.00, 2007, Toronto, ON Canada: Advanced Chemistry Development Inc, available at: http://www.acdlabs.com

[b]Istituto di Ricerche Farmacologiche Mario Negri, Milano, Italy, (2010) available at: http://www.insilico.eu/coral/

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Experimental and calculated pEC50 values for split 1, 2, and 3.

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.