On the agreement of external validation parameters for linear regression QSAR models

Nicola Chirico, Ester Papa and Paola Gramatica

QSAR Research Unit in Environmental Chemistry and Ecotoxicology, Department of Structural and Functional Biology, University of Insubria, Varese, Italy.

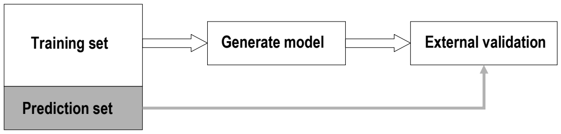
E-mail: nicola.chirico@uninsubria.it; paola.gramatica@uninsubria.it - web: www.qsar.it

ABSTRACT

The evaluation of linear regression QSAR models performances, both in fitting and external prediction, is of pivotal importance [1][2]. While leave-one-out (LOO) Q^2 internal validation technique (cross-validation) is well established, different external validation parameters have been proposed in the last decade: Q^2_{F1} (Shi) [3] , Q^2_{F2} (Schüürmann) [4], Q^2_{F3} (Consonni) [5][6], r^2m (Roy) [7] and the Tropsha-Golbraikh [8] method. These parameters usually are in accordance, making one confident of a model predictivity, but doubts arise when they give contradictory results. In these cases the QSAR model developer should understand which one of the aforementioned parameters is "the best". However this is not an easy task, mainly because no one of these parameters could be considered "the best" in every situation. We are thus looking for a simpler method to evaluate the external predictivity of the models, independently on the set composition. In our opinion, the simplest method consists in the quantification of the similarity among the experimental data of external test set versus the corresponding values calculated by the model.

In this study the concordance correlation coefficient [9] has been used as a reference and we have evaluated the number of contradictory and agreeing results on validation parameters by means of 210.000 simulated datasets. A wide range of possible scenarios has been generated and, concerning the more realistic ones, 95% of agreement has been found among the concordance correlation coefficient and all the aforementioned validation parameters together. Our proposed coefficient is the most precautionary among those analyzed. We have verified that disagreements among results is related to two possible situations: a) the external data points are well predicted (good matching), while at least one of the validation parameters rejects the model (rare), b) the matching is not good and one or more validation parameters accept the model (relatively common). The second alternative is more dangerous for QSAR models, thus a deeper analysis of the results is suggested. Our method, verified also on real models, has been proposed as a tool to be used in addition, or even in alternative, to the aforementioned external validation parameters to find out this kind of critical models with doubtful predictivity.

EXTERNAL VALIDATION PARAMETERS



External validation is basically based on two techniques:

- Q^2 formulas
- Experimental vs predicted responses

Q^2 formulas

$$Q_{F1}^{2}=1-\frac{\sum_{i=1}^{n_{EXT}}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i-\bar{y}_{TR})^2}\quad [3]$$

$$Q_{F2}^{2}=1-\frac{\sum_{i=1}^{n_{EXT}}(\hat{y}_i-y_i)^2}{\sum_{i=1}^{n_{EXT}}(y_i-\bar{y}_{EXT})^2}\quad [4]$$

$$Q_{F3}^{2}=1-\frac{\left[\sum_{i=1}^{n_{EXT}}(\hat{y}_i-y_i)^2\right]/n_{EXT}}{\left[\sum_{i=1}^{n_{TR}}(y_i-\bar{y}_{TR})^2\right]/n_{TR}}\quad [5][6]$$

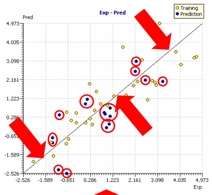
Experimental vs predicted responses

GOLBRAIKH AND TROPSHA METHOD [8]

- R^2 and R^2_0 (origin forced)
- Angular coefficients
- Closeness: $(R^2 - R^2_0) / R^2$

Calculated for both axes dispositions (predicted values vs. experimental / experimental vs. predicted)

$$r^2m = r^2\left(1-\sqrt{r^2-r_0^2}\right)\quad [7]$$



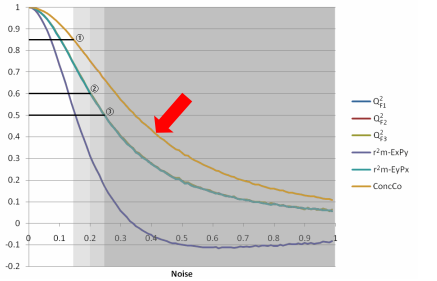
WE PROPOSE

Concordance correlation coefficient [9]

$$\hat{\rho}_c=\frac{2\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2+\sum_{i=1}^{n}(y_i-\bar{y})^2+n(\bar{x}-\bar{y})^2}$$

It is similar to the correlation coefficient but takes into account the diagonal (perfect match)

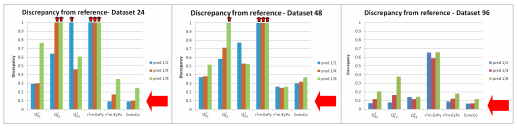
GENERAL PERFORMANCES USING BIG SIMULATED DATASETS



For every value of noise, 10^6 simulated datasets are generated. The concordance correlation coefficient is the most restrictive.

Thresholds are in the encircled numbers: 1) concordance correlation coefficient and r^2m-ExPy (experimental data on the abscissa and predicted values on the ordinate) – rejection region starting from light gray, 2) Golbraikh and Tropsha method, Q^2_{F1}, Q^2_{F2} , Q^2_{F3} – rejection region starting from middle gray, 3) r^2m-EyPx (predicted data on the abscissa and experimental values on the ordinate) – rejection region in dark gray.

DISCREPANCY FROM REFERENCE – DATASETS OF REALISTIC SIZE



Considering the values obtained from the simulated models, the discrepancy from reference values of the concordance correlation coefficient is small, expecially for the smallest dataset.

CONCLUSIONS

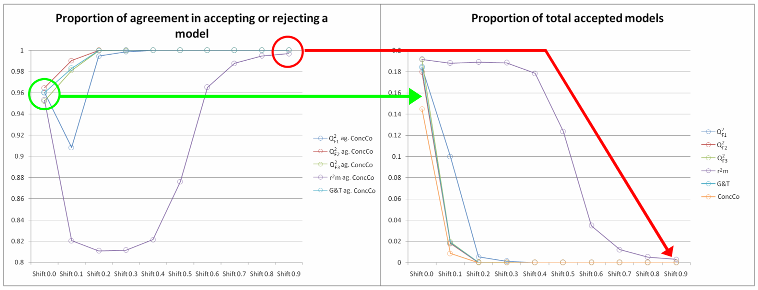
- ✓ The here proposed concordance correlation coefficient (ConcCo) is conceptually simple, being similar to a correlation coefficient, and proved to be the the most restrictive parameter in accepting models using big simulated datasets.
- ✓ ConcCo is in good agreement (96%) with the other parameters when datasets of realistic sizes are simulated. In the remaining situations, when the parameters are discordant, ConcCo is the most restrictive in almost all the cases.
- ✓ ConcCo is the most reliable (stable) parameter in the studied real datasets. Therefore, when the validation parameters disagree, ConcCo helps to make a decision whether a model should be accepted or not as predictive.
- ✓ **Paper submitted to J. Chem. Inf. Mod.**

MATERIAL AND METHODS

Datasets are generated at random, following a gaussian distribution, using a custom simulation software. Datasets sizes span from 10^6 elements for the general parameter performances to 24-1536 for the realistic ones (210.000 simulated datasets). Prediction set proportions for the realistic sizes are: 1/2, 1/4 and 1/8. Parameter performances are calculated over different level of noise in both the training and prediction set responses and different levels of systematic shifts in the prediction set responses.

Real datasets have been also taken from literature [10-14] to compare the different validation parameters in real QSAR scenarios.

AGREEMENT AMONG THE PARAMETERS - DATASETS OF REALISTIC SIZE



The external validation parameters agree 96% (green circle) of the times with the concordance correlation coefficient (ConcCo) when no systematic shift is added (more realistic situation).

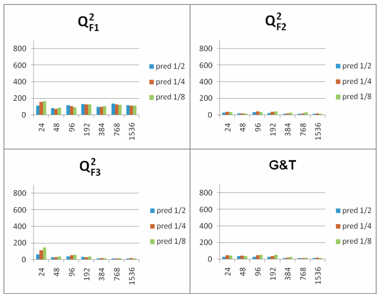
96% agreement (green circle) with the concordance correlation coefficient is found mainly in accepting models.

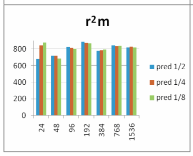
100% agreement (red circle) is reached when all the parameters reject all the models (shift 0.9)..

The parameter ConcCo is the one accepting the least number of models – most cautelative

RESTRICTIVENESS COMPARISON – DATASETS OF REALISTIC SIZE

ConcCo rejects models when the other parameteres accept them





ConcCo = concordance correlation coefficient

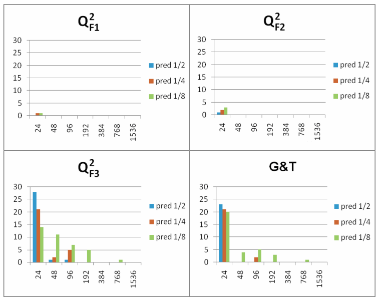
Ordinate: number of models

Abscissa: dataset size

pred 1/n: prediction set proportion

It is relatively common that ConcCo rejects "critical" models accepted by one ore more of the other parameters.

ConcCo accepts models when the other parameteres reject them



It is relatively rare that ConcCo accepts "critical" models rejected by one ore more of the other parameters.

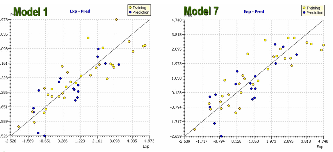
Overall, the here proposed concordance correlation coefficient (ConcCo) proved to be more restrictive in accepting models than the other parameters.

REAL QSAR SCENARIOS

Nitro-PAH mutagenicity models with discordant external validation parameter values

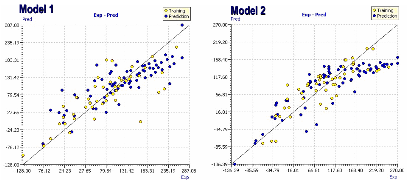
ID	Variables	R^2	Q^2_{LOO}	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	ConcCo	r^2m-ExPy	r^2m-EyPx	G&T
1	PW2 IC1	0.81	0.77	0.60	0.53	0.73	0.79	0.63	0.53	accepted
2	BELe8 HATS4u	0.81	0.76	0.48	0.38	0.65	0.73	0.52	0.46	accepted
3	VED2 R6u+	0.79	0.76	0.27	0.14	0.51	0.58	0.26	0.28	rejected
4	HATS3u R3v	0.80	0.76	0.00	0.00	0.00	0.39	0.17	0.11	rejected
5	BELe8 R4u+	0.80	0.75	0.50	0.42	0.67	0.74	0.55	0.46	accepted
6	SIC2 BEHm8	0.79	0.75	0.61	0.55	0.74	0.8	0.6	0.51	accepted
7	SIC2 BELv5	0.79	0.75	0.58	0.51	0.72	0.78	0.58	0.48	Rejected

ConcCo = concordance correlation coefficient, r^2m-ExPy = experimental values on the abscissa axis, r^2m-EyPx = experimental values on the ordinate axis, G&T = Golbraikh and Tropsha method.



Graphs are comparable, ConcCo has only a small variation and rejects both the models. Larger variations are observed in some of the other validation parameters.

PFCs boiling point models with discordant external validation parameter values



ID	Q^2_{F1}	Q^2_{F2}	Q^2_{F3}	ConcCo	r^2m-ExPy	r^2m-EyPx	G&T
1	0.74	0.69	0.67	0.79	0.40	0.63	accepted
2	0.69	0.68	0.47	0.79	0.53	0.71	reject

ConcCo = concordance correlation coefficient, r^2m-ExPy = experimental values on the abscissa axis, r^2m-EyPx = experimental values on the ordinate axis, G&T = Golbraikh and Tropsha method.

The same results as above are observed for ConcCo while almost all of the other validation parameters have larger variations.

REFERENCES

[1] Tropsha et al. The importance of Being Earnest: Validation in the Absolute Essential for Successful Application and Interpretation of QSPR Models. QSAR Comb. Sci. 2003, 22, 69-76

[2] Gramatica. Principles of QSAR models validation: internal and external. QSAR Comb. Sci. 2007, 5, 694-701

[3] Shi et al. QSAR Models Using a Large Diverse Set of Estrogens. J. Chem. Inf. Comput. Sci. 2001, 41, 186-195.

[4] Schüürmann et al. External Validation and Prediction Employing the Predictive Squared Correlation Coefficients Test Set Activity Mean vs Training Set Activity Mean. J. Chem. Inf. Model. 2008, 48, 2140–2145

[5] Consonni et al. Comments on the Definition of the Q^2 Parameter for QSAR Validation. J. Chem. Inf. Model. 2009, 49, 1669-1678

[6] Consonni et al. Evaluation of model predictive ability by external validation techniques. J. Chemom. 2010, 24, 194–201

[7] Roy. On some aspects of validation of predictive quantitative structure-activity relationship models. Expert Opinion on Drug Discovery, 2007, 2, 1567-1577

[8] Golbraikh and Tropsha. Beware of q^2. J. Mol. Graph. Model. 2002, 20, 269-276.

[9] Lin. A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics 1989, 45, 255-268

[10] ENV/JM/MONO(2004)24. http://appli1.oecd.org/olis/2004doc.nsf/linkto/env-jmmono, p. 99. (accessed April 13, 2011)

[11] Gramatica et al. Approaches for externally validated QSAR modeling of Nitrated Polycyclic Aromatic Hydrocarbon mutagenicity. SAR QSAR Environ. Res. 2007, 18 , 169-178

[12] http://chem.sis.nlm.nih.gov/chemidplus/ (accessed April 13, 2011)

[13] Hendricks J. O. Industrial fluoro-chemicals. Ind. Eng. Chem. 1953, 45, 99-105

[14] Bhhatarai et al. CADASTER QSPR Models for Predictions of Melting and Boiling Points of Perfluorinated Chemicals. Mol. Inf. 2011, 30, Volume: 30, 189-204

This work was supported by the
European Union through the project CADASTER FP7-ENV-2007-1-212668