Linear and Nonlinear Methods in Modeling the Aqueous Solubility of Organic Compounds

Cornel Catana*

CADD, Pfizer Global Research and Development, Ann Arbor Laboratories, 2800 Plymouth Road, Ann Arbor, Michigan 48105

Hua Gao

CADD, Pfizer Global Research and Development, Groton Laboratories, Eastern Point Road, Groton, Connecticut 06340

Christian Orrenius and Pieter F. W. Stouten

Computational Sciences, Nerviano Medical Science, Viale Pasteur 10, 20014 Nerviano (MI), Italy

Received June 23, 2004

Solubility data for 930 diverse compounds have been analyzed using linear Partial Least Square (PLS) and nonlinear PLS methods, Continuum Regression (CR), and Neural Networks (NN). 1D and 2D descriptors from MOE package in combination with E-state or ISIS keys have been used. The best model was obtained using linear PLS for a combination between 22 MOE descriptors and 65 ISIS keys. It has a correlation coefficient (r^2) of 0.935 and a root-mean-square error (RMSE) of 0.468 log molar solubility (log S_w). The model validated on a test set of 177 compounds not included in the training set has r^2 0.911 and RMSE 0.475 log S_w. The descriptors were ranked according to their importance, and at the top of the list have been found the 22 MOE descriptors. The CR model produced results as good as PLS, and because of the way in which cross-validation has been done it is expected to be a valuable tool in prediction besides PLS model. The statistics obtained using nonlinear methods did not surpass those got with linear ones. The good statistic obtained for linear PLS and CR recommends these models to be used in prediction when it is difficult or impossible to make experimental measurements, for virtual screening, combinatorial library design, and efficient leads optimization.

1. INTRODUCTION

The implication of aqueous solubility as an important factor in absorption and consequently in overall bioavailability of drugs has led to an intense activity applying statistical modeling for the prediction of solubility based solely on chemical structure. The tedious experimental procedure of solubility measurements and the need to estimate solubility for structures of compounds that do not physically exist have added incentive to the search for truly alternative high-throughput methods. Consequently, the direct address of well-known obstacles of physical nature in solubility predictions such as solid-state forces/crystal lattice energies and ionization states of molecules are avoided (melting points for nonexisting compounds are difficult to determine) with the hope that such phenomena would be taken into account indirectly during training. Important contributions to the field of solubility predictions where these assumptions were considered unacceptable are exemplified by the work of Yalkowsky[1] and Abraham.[2] However, indications are that the former procedure could be sufficient for certain applications and that experimental error in determinations often overrides any other source of error in limiting final overall performance.

In principle, an exhaustive screening of statistical methods in order to optimize performance would involve linear, nonlinear, and no feature/variable/descriptor selection in combination with linear and nonlinear model construction. Thus, a total of six different results could be obtained, and the best performing approach should be implemented/applied. Because the experimental solubility data used in this study cover a large range and nonlinear correlations are likely, we report results obtained by three principally different approaches: linear and nonlinear PLS, CR, and NN.

2. METHODS

Data Sets. The data set was almost identical (training set identical, test set slightly different) with that used in QSPR Analysis of Aqueous Solubility for a Diverse Set of Organic Compounds, Study Report A0082954-Pharmacia, 2000, authors H. Gao, A. Vulpetti, and Pil H. Lee. In summary, the aqueous solubility data set used included 800 compounds, the majority (473) from AquaSol[3] database, and 307 in-house compounds. The AquaSol database contains large amounts of solubility records extracted from a number of scientific references. It contains sometimes several experimental values for the same compound, and in this case the highest solubility measurement was taken into consideration. The database covers a variety of compounds including pharmaceuticals, pollutants, nutrients, herbicides, pesticides, and agricultural

* Corresponding author phone: (734)622-4479; fax: (734)622-2782; e-mail: cornel.catana@pfizer.com.

10.1021/ci049797u CCC: $30.25 © 2005 American Chemical Society
Published on Web 11/24/2004

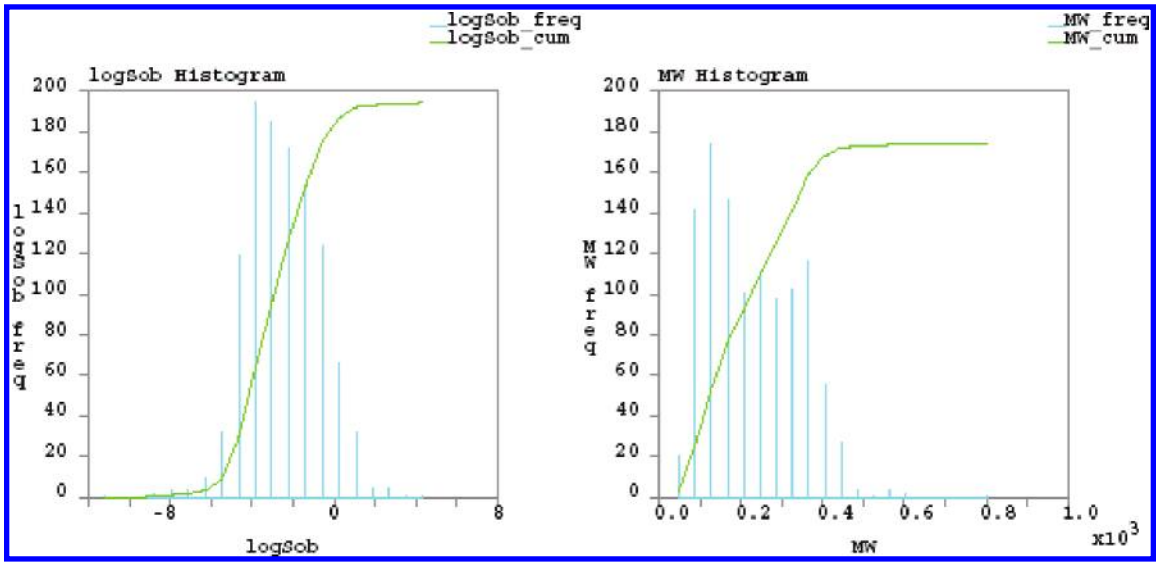

Figure 1. The distribution of the solubility data and MW for 1107 compounds analyzed.

and industrial compounds. Most of the compounds in AquaSol are simple, nonfunctional chemicals or environmentally interesting compounds. To increase the representation of drug-like compounds in the set, data collected from the literature for pharmaceuticals and drug-like molecules[4−13] and our in-house solubility measurements were included. All solubility data were converted to the log of molar solubility (moles of solutes liter^{-1}). The aqueous solubility values used in this study ranges from −11.62 to 4.75. The distribution of the solubility data analyzed and MW is plotted in Figure 1. Diversity analysis[14] was used to select a diverse training set of 930 compounds; the remainder, 177 compounds, made up the test set. For the initial set of 1107 compounds, ISIS keys descriptors (166) have been calculated, and the Monte Carlo algorithm was used to optimize the diversity function (MaxMin)

$$\text{MaxMin} = \max\{\min\{D_{ij}^{\,2}\}\}$$

where D_{ij} is the distance in the Euclidean space; the diverse set was extracted based on the Tanimoto similarity coefficient. This procedure usually increases the chances to get on the test set a similar statistics (r^2, RMSE, etc.) with that on the training set. As a rule of thumb all the models were developed on the training set; the test set (unseen set in modeling process) was used only for prediction.

Descriptor Selection. 1D and 2D descriptors from MOE (Molecular Operating Environment)[15] were used. In this way speed in calculations is ensured, and the issue of how to generate and select 3D conformations when calculating the descriptors is eliminated.

In MOE there are quite a few descriptors based on the concept of creating a descriptor value for a specific range $[u,v)$ of the property values P; for example a descriptor value will equal the sum of the atomic van der Waals (VSA) contributions of each atom i with P_i in $[u,v)$. More precisely the quantity $P\_\text{VSA}(u,v)$ is defined like

$$P\_\text{VSA}(u,v) = \sum_i V_i \delta(P_i \in [u,v))$$

Here V_i is the atomic contribution of atom i to the VSA of the molecule. Further, a set of n descriptors associated with the property P can be defined as follows:

$$P\_\text{VSA}_k = \sum_i V_i \delta(P_i \in [a_{k-1}, a_k)), \quad k = 1,2,...n$$

Here $a_0 < a_k < a_n$ are interval boundaries such that $[a_0, a_n]$ includes all values of P_i in any molecule. Each VSA-type descriptor can be characterized as the amount of surface area with P in a certain range. For example SlogP_VSA6 equals the surface area of those atoms (within a molecule) with logP values in the interval (0.20, 0.25]. More details about these descriptors and how the interval boundaries are determined have been described in ref 16.

As a starting point for our studies, all the VSA-type descriptors corresponding to a subdivision of the molecular surface for the following properties—refractivity, octanol/water partition, and partial charge—were used, in addition to all topological and 1D descriptors a total of 146.

Genetic algorithm (GA), included in the PLS-Toolbox from MATLAB, was used for variables selection (based on the training set) with PLS as a regression method, and the cross-validation procedure was performed randomly over 10 subsets in 5 iterations. The population size was 128, maximum generation was set to 100, and double crossover and a mutation rate of 0.005 were used. In addition QuaSAR-Contingency[15] module from MOE was used to identify (it extracts linear and nonlinear descriptors) and to rank the most important descriptors. QuaSAR-Contingency module includes a contingency analysis and an uncertainty coefficient that attempt to measure the degree to which two random variables are dependent. The 2 lists of descriptors one provided by GA and the other by QuaSAR-Contingency analysis have been overlapped, and the result (CL − consensus list) included 22 descriptors; more, these descriptors are ranked almost in the top of the contingency list. The 22 descriptors deemed as important in their correlation with experimental solubility are presented below.

Hydrophobicity Descriptors. MOE implements logP(o/w) calculated by the recent method of Wildman and Crippen's;[17] its correlation with logSob, for the training set, is pretty high $r^2 = 0.3$.

In addition to logP, SlogP_VSA6 − SlogP_VSA9 (in the interval 0.2, 0.25, 0.3, 0.4, infinite) were selected as important by the procedure described above.

Refractivity Descriptors. The molar refractivity (SMR) incorporates both the size and the polarizability of a molecule. Taking into account that solvation (one of the important steps of solubility process) could be considered as generating cavities occupied by the solute molecules in the solvent, we thought including such a descriptor would be beneficial. Its subdivisions SMR_VSA5 − SMR_VSA7, which were selected in addition, are in the interval boundaries (0.44, 0.485, 0.56, infinite). The sum of atomic polarizabilites (apol) was also found to correlate with solubility.

Atomic Partial Charge Descriptors. These capture the electrostatic interactions. MOE uses the Partial Equalization of Orbital Electronegativities (PEOE) method of Gasteiger[18] to calculate partial charges, which is based upon the iterative equalization of atomic orbital electronegativities.

CL has found as important descriptors those that bear negative charge, like PEOE_VSA-4 − PEOE_VSA-6 (the interval boundaries being −0.20, −0.15, −0.10, −0.05), as those with positive charge like PEOE_VSA+4 − PEOE_VSA+6 (the interval boundaries being 0.2, 0.25, 0.30, infinite).

In addition, total hydrophobic vdw surface area, PEOE_VSA_HYD (this is the sum of the v_i -van der Waals surface area of atom *i*- such that $|q_i|$ is less than or equal to 0.2), and total positive van der Waals surface area, PEOE_VSA_POS (this is the sum of the v_i such that q_i is nonnegative), were selected.

Topological Descriptors. CL included four descriptors, topological polar surface area (TPSA) being the most important; it is a measure of polar surface area (hydrophilicity) for 2D structures. Even if molecular flexibility (KierFlex), bond connectivity index of order 2 (chi2), and vertex index magnitude (VdistMa) are less meaningful in terms of their interpretability to the physical process of solubility, they were found to make an important contribution to the model.

The 22 descriptors were included in a multilinear regression (MLR) equation, and the result obtained was good in terms of statistics: $r^2 = 0.9$ and RMSE $= 0.57$ log S_w. Further analysis showed that these descriptors are uncorrelated. Applying this equation to the test set we got RMSE $= 0.55$ log S_w.

Does the addition of other 2D descriptors from MOE impact the statistics? After we removed the 22 descriptors from the contingency list we found at the top electrotopological state indices, E-State (we implemented them in MOE), already used in studies of this kind;[19,20] we observed a slight enhancement of the MLR model when the list of descriptors included them in addition to the 22. In a later stage of our studies we reached the conclusion that by replacing E-State descriptors with ISIS keys the statistics enhances slightly and as a consequence, respecting the chronology of the events, these results are reported separately.

Focusing on E-State descriptors we removed those which did not map in any of the compounds from the training set (frequency zero); in this way 40 E-State descriptors have been added to the previous 22 mentioned before. The unified set of descriptors, 62, has been analyzed by different techniques.

3. TECHNIQUES

PLS-Toolbox[21] was used to carry out the multivariate analysis and Statistica, Neural Networks[22] to develop nonlinear models.

Linear PLS Modeling. The linear PLS model was cross-validated and constructed interactively using the 'modlgui' function. The data were auto-scaled. Of the two PLS methods implemented in the toolbox, NIPALS[23] and SIMPLS,[24] the latter was used, and cross-validation was done using Venetian blinds leaving out every tenth sample as the test set. The root-mean-square error of calibration (RMSEC), identical with RMSE, evaluates the fit of the model to the calibration (training) data and is defined as

$$\mathrm{RMSEC} = \mathrm{sqrt}(\sum(y_{\mathrm{ipred}} - y_i)^2/n)$$

Here y_{ipred} are the values of the predicted variable when all samples are included in model formation, and n is the number of calibration samples. y_i are the experimental values. This is in contrast to the root-mean-square error of cross-validation (RMSECV $\equiv$ RMSECV$_k$ where k is the number of latent variables for which we get the minimum) which is a measure of the model's ability to predict new samples. The RMSECV$_k$ is defined as in the above equation, where y_{ipred} are predictions for samples not included in model formation. RMSECV$_k$ is related to the PRESS$_k$ value through the number of latent variables, LVs (known also as components, factors), included in the model

$$\mathrm{RMSECV}_k = \mathrm{sqrt}(\mathrm{PRESS}_k/n)$$

Here PRESS$_k$ is the sum of squares prediction error for the model that includes k LV. It is possible to calculate a root-mean-square error of prediction (RMSEP) when the model is applied to unseen data provided that the reference values for the new data are known. The formula is the same as RMSEC except that the estimates are based on a previous developed model.

Nonlinear PLS Modeling. PLS with polynomial and spline inner relations are implemented in PLS-Toolbox; the nonlinear term refers to the inner relation. For example, a polynomial cubic relation is defined as

$$u = b_0 + b_1 t + b_2 t^2 + b_3 t^3$$

where u and t are the Y respectively X-scores of the LV.

In a polynomial fit, the user has the possibility to specify the polynomial order for the PLS inner relation.[25] High values in polynomial order and/or LV can lead to severe "overfitting".

The second nonlinear variant of PLS uses a spline to fit the inner relation[26] and is named SPL-PLS. Besides the maximum number of latent variables and the degree of the spline, the number of knots used influences the final result.

Continuum Regression (CR). The regression techniques of MLR, PCR (Principal Component Regression), and PLS can all be unified under one approach which is referred to as continuum regression.[27] The central idea behind continuum

regression is that the PLS method captures covariance between the X- and Y-blocks. This can be thought of as an attempt to balance the two tasks of providing a reduced order description of the input data block and correlating the input data to the output data. At opposite extremes of this trade off are PCR, which starts with a model that describes variance in the X-block correlating it piece by piece to the Y-block and MLR (seeks only to correlate the X- and Y-blocks without regard to the input block structure). The conventional PLS method tries to do both. When using CR, cross-validation must be done to determine both the number of LV and the techniques within the continuum that produces the optimum model. The first step in this method is to perform a singular value decomposition (SVD) on the X-block. When the singular values are taken to a power m greater than 1 the CR goes toward PCR ($m=\infty$). When m goes toward 0, CR approaches MLR ($m=0$). For $m = 1$, CR is identical to PLS.[28]

Neural Networks (NN). Selecting an appropriate network type and architecture (number of hidden unit and parameters of training algorithms)[29−31] can be time-consuming and a unrewarding activity, involving a large number of heuristic experiments. 'Statistica Neural Networks' Automatic Network Designer[22] (part of Statistica) performs these tedious tasks and selects the best network architecture and size. Using this functionality, 2 types of networks were investigated: multilayer perceptron (MLP) and linear networks.

MLP[31,32] is the most popular network architecture in use today. The number of input and output units is defined by the problem; the number of hidden layers and units to use has to be determined from case to case (we used in this study 1 and 2 hidden layers). The most common error function is the squared sum error, where the individual errors of output units on each case are squared and summed together (this is our case). The activation function is determined automatically by Automatic Network Designer from a pool of functions such as logistic, hyperbolic, and exponential among others.

The linear network provides a good benchmark against which to compare the performance of your neural network. It is quite possible that, problems regarded to be highly complex, can actually be solved as well by linear techniques.

A linear network has only two layers: an input and output layer, the latter having a linear activation function. The linear network was trained using the pseudoinverse technique,[32] a fast method that guarantees to find the optimal solution.

4. RESULTS AND DISCUSSION

Linear and Nonlinear PLS. PLS (linear) shows a model with 20 latent variables. The variance captured by X-block and Y-block is 66%, respectively 93%; RMSEC = 0.497, RMSECV = 0.570 log S_w. The analysis of loading variables pointed toward the following descriptors in the first 6 LV: TPSA, SMR, apol, logP(o/w), PEOE_VSA_HYD, SlogP_VSA6, VDistMa, S_dCH2, S_dssC. In the prediction (validation) phase RMSEP = 0.495 log S_w. This corresponds to a correlation coefficient $r^2 = 0.905$. The observed and calculated solubility data of the training and test sets are plotted in Figures 2 and 3, respectively.

Polynomials of an order between 2 and 5 were used for polynomial PLS, the number of LV being 20 (obtained as optimum by cross-validation in linear PLS).

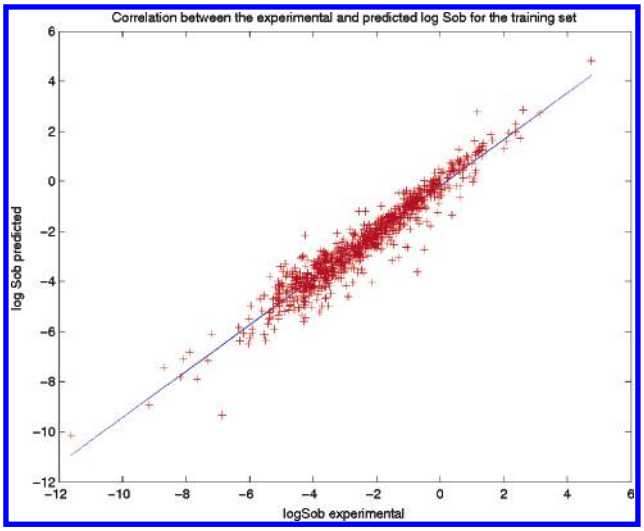

Figure 2. Correlation between the experimental and predicted log S for training set.

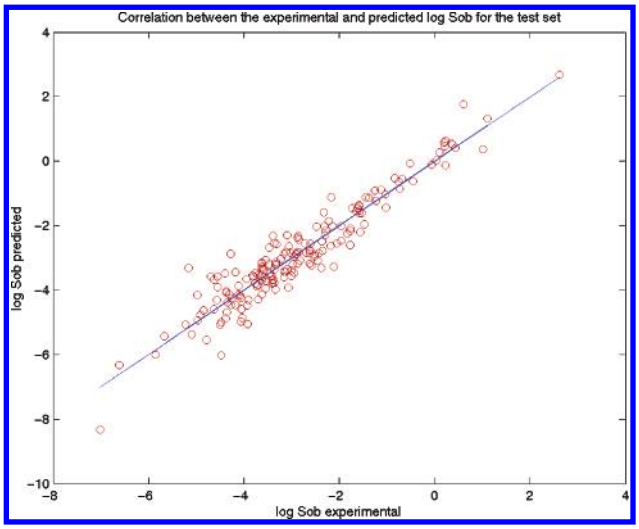

Figure 3. Correlation between the experimental and predicted log S for test set.

Table 1. Influence of the Polynomial Order on Explained Variance of the Training Set and on the RMSEP of the Test Set

polynomial order	% variance captured by X-block	% variance captured by Y-block	RMSEC	RMSEP
1	66.00	93.00	0.497	0.570
2	66.46	92.43	0.495	0.535
3	67.50	92.77	0.491	0.486
4	67.11	93.03	0.483	0.499
5	67.38	93.04	0.495	0.537

To avoid the "over-fitting" with respect to polynomial order, the RMSEP for the test set was calculated. The best compromise (see Table 1) is reached with a cubic polynomial: RMSEC = 0.491 log S_w, RMSEP = 0.486 log S_w. These results are a slightly better than what was obtained with linear PLS.

For SPL-PLS the optimum number of LV found by linear PLS cannot be extrapolated. Our experience showed that the results worsen in prediction when the number of LV is above 10. The polynomial degree also influences the results. A script in MATLAB was written to find the optimum combination between the number of LV, the polynomial degree, and the number of knots (see Table 2). Only the best

Table 2. Influence of the Polynomial Order, LV, and Number of Knots on Explained Variance of the Training Set and RMSEP of the Test Set

knots	LV	polynomial order	% variance captured by X-block	% variance captured by Y-block	RMSEP
5	5	2	35.58	90.00	0.553
6	4	2	38.35	89.96	0.554
6	5	2	35.40	89.85	0.555
4	5	2	35.11	90.16	0.560
2	8	3	43.29	90.30	0.558
2	8	3	41.33	90.09	0.562
4	4	3	30.23	88.89	0.563
4	5	3	33.64	89.13	0.568

results in term of RMSEP are presented. The number of LV was varied between 3 and 10, the number of knots used 1 to 8 for the polynomial order 2 and 3. Practically no variation of RMSEP values with polynomial order 2 or 3 was observed. The best results were obtained for 5 knots and 5 LV for a polynomial order of 2. The results are inferior to those obtained using polynomial PLS indicating that a spline correlation is unlikely.

Investigation of solubility with nonlinear PLS revealed that only cubic polynomial PLS behaves slightly better than PLS and further tests (prediction on unseen compounds) indicated that the model is over trained.

Continuum Regression. The power (m) was varied in the range 0.1−10 (step size 0.05). As in PLS, the cross-validation was done using venetian blinds leaving out every tenth sample as the test set. Random reordering of the data set was performed 5 times. The results are sensitive to this parameter. Multiple runs were performed in order to come with an optimized solution. First, CR models were calculated using a maximum of 20 LV. In Figure 4, a representation of the predictive residual error sum of squares (PRESS) matrix as a function of the number of the LV and power (m) is shown. CR produced a model situated between PLS (m=1) and MLR (m=0), and the optimum values calculated were LV = 6, m = 0.398. For these parameters, which indicate linearity in our data set, RMSEC is 0.497 log S_w. Applying this model to the test set a RMSEP = 0.497 log S_w was obtained (results almost identical with those obtained by linear PLS).

We looked also at the CR prediction error surface, that means all the CR models were applied to the test set. The smallest RMSEP = 0.488 log S_w was obtained for LV = 9, m = 0.562. This optimum is close to the solution found through cross-validation.

Table 3. Network Architectures Found To Model Best the Solubility Data

type	inputs	hidden	RMSE train	RMSE verify	RMSE test	r^2 train	r^2 verify	r^2 test
MLP	41	12	0.584	0.529	0.608	0.897	0.923	0.846
linear	60		0.483	0.532	0.501	0.930	0.921	0.903

Neural Networks. Running the "Automatic Network Designer", the MLP with one hidden layer and linear networks proved to best model our data (see Table 3).

Given the fact NN is very sensitive to the number of input variables it is recommended to run Automatic Network Designer in combination with GA (from STATISTICA). Following this procedure the initial number of 62 variables was further reduced to 41 for MLP respectively to 60 for linear NN. In MLP the hidden layer had 12 neurons (nodes), Levenberg−Marquardt algorithm was used for training, and the activation function was logistic.

Because NN uses in addition the verification set, this is the case when early stopping technique is used to prevent over-fitting, the initial training set (930 compounds) has been divided in 800 compounds for training and 130 for verification set. The test set remained unaltered. The set of 130

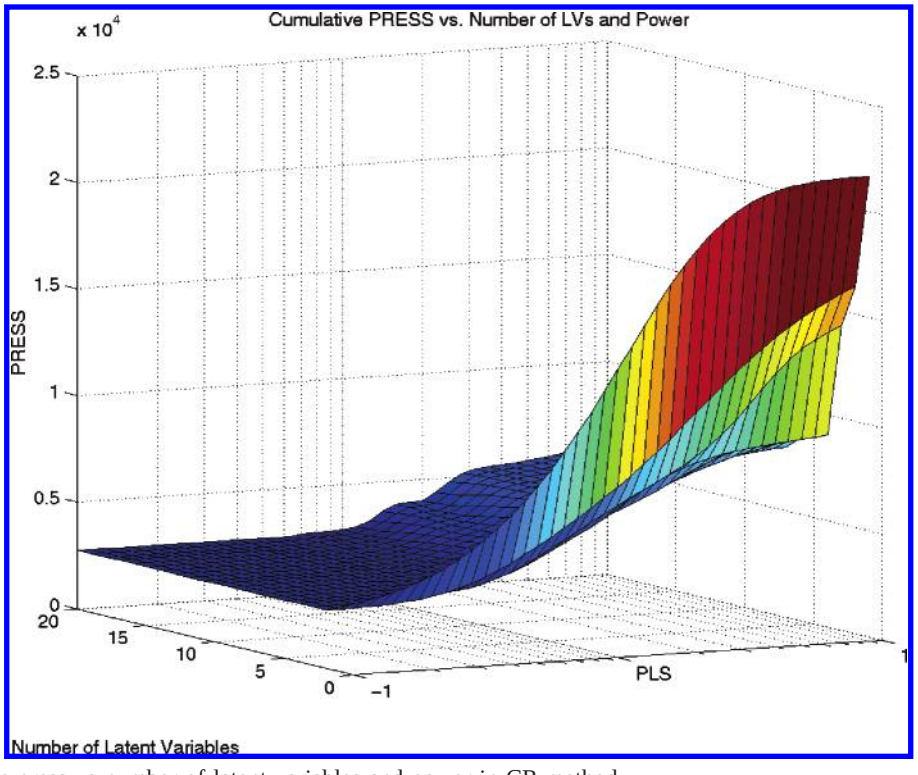

Figure 4. Cumulative press vs number of latent variables and power in CR method.

compounds was chosen by shuffle technique. The RMSE (train, verify, and test) values are reported in Table 3.

More, other nonlinear algorithms (see ALOGPS program available for free download from VCCLAB site http://www.vcclab.org) were used to predict the test set to avoid any bias toward our conclusions. The LIBRARY of ALOGPS allows instant retraining of the program for new data (in this case our training set) using Associative Neural Networks (ASNN) and, based on this, the prediction for the test set. The results we got, RMSE test = 0.576, r^2 test = 0.855, places the performances of ASNN (which uses E-state indices) between MLP and linear NN.

If we take into consideration the fact that NN models are difficult to interpret we could conclude that for the current set of compounds and their assigned descriptors NN do not bring any improvement.

ISIS Key Descriptors. E-State descriptors have an inconvenience in our treatment; when they are not present in a compound, the value assigned is zero and not considered as missing. Taking this into account, we could replace E-state descriptors with a value equal to "1" when they are present and "0" when they are not. Applying this procedure we got models almost as good as those reported above so we decided to replace E-state descriptors with ISIS public keys (166 descriptors), to see how this will affect our results.

GA from PLS-Toolbox (MATLAB) in combination with QuaSAR-Contingency analysis was used for ISIS keys selection; the procedure was identical with that reported in the identification of the 22 MOE descriptors. The 65 ISIS keys found as important for solubility have been unified with the 22 MOE descriptors, and the new set of 87 descriptors was used as input for the different techniques reported below. Other scenarios would have been possible, for example to unify MOE descriptors with ISIS key or to unify E-Sate with ISIS key and MOE descriptors and then to apply GA and QuaSAR-Contingency to identify the most important one.

First we applied linear PLS; cross-validation pointed toward 40 latent variables. The variance captured by X-block and Y-block is 88.3%, respectively 93.5%; RMSEC = 0.468, RMSECV = 0.557 log S_w. In the prediction phase RMSEP = 0.475 log S_w; this corresponds to correlation coefficient r^2 = 0.911. Besides the 22 descriptors, PLS found the following ISIS keys as important:

- ACH_2QH (key 82) where Q stands for any atom except H and C and A stands for any atom except H;
- Aromatic ring > 1 (key 125);
- N heterocycle (key 121);
- QSQ (key 58) and OSO (key 55);
- CH_3CH_2A (key 114), etc.

Depending on their sign these ISIS keys correlate positively or negatively with the solubility.

Because the linear PLS pointed for linearity (40 LV) we skipped over the nonlinear methods and applied directly CR, which has found a model situated nearby MLR (m=0), the optimum values calculated being: LV = 2, m = 0.112. For these parameters, which indicate high linearity in our data set, RMSEC = 0.472 log S_w; for the test set we got a RMSEP = 0.477 log S_w result almost identical with that obtained by linear PLS.

In term of NN we got the best results with a linear one; the number of descriptors has been reduced from 87 to 58 and the RMSE on Train/Verify/Test = 0.485/0.504/0.487 log S_w. The results are not superior to those obtained by linear PLS or CR so our general conclusion is that the last 2 models could provide the best results in solubility prediction for new compounds because of their statistics and the way in which cross-validation has been done. In addition we conclude that for the given set of compounds ISIS keys can replace with success the E-state descriptors. Studies done on other internal sets of compounds showed us that the combination between MOE descriptors (the type we mentioned in this paper), E-state, and ISIS keys provides good results in modeling the solubility. In terms of their importance QuaSAR-Contingency analysis ranked the MOE descriptors at the top of the list followed by a mixture of E-state descriptors and ISIS keys.

5. CONCLUSIONS

Linear and nonlinear methods have been carried out to determine which of them model best the relationship between experimental solubility and a set of 1D and 2D descriptors. The combination between 22 descriptors from MOE with a subset of 65 ISIS keys, used in a linear PLS method, provided the best statistics for the given set of compounds. The variance captured by X-block and Y-block is 88.3% and 93.5%, respectively, RMSEC = 0.468 log S_w and RMSECV = 0.557 log S_w. The validation (prediction) was carried out on 177 compounds not included in the training set and a RMSEP = 0.475 log S_w was calculated; this corresponds to a correlation coefficient r^2 = 0.911. A similar statistic was obtained when we used as method CR. On other test sets both methods will be used in parallel to see which will give better results in prediction.

The nonlinear methods we developed are slightly inferior to linear ones. In our opinion there are 2 possible explanations: (a) there are no significant nonlinear dependencies between the molecular descriptors used and the aqueous solubility analyzed or (b) being on the limit of experimental error the nonlinear methods cannot behave better than the linear ones.

The descriptors were ranked according to their importance using QuaSAR-Contingency analysis; MOE descriptors (of the type mentioned in this paper) are at the top of the list followed by a mixture of E-state and ISIS keys.

Since these models are based on a set of calculated 1 and 2D descriptors (which ensure a high speed) they might be used for assessment of aqueous solubility in the design of virtual combinatorial libraries, for efficient leads optimization, and in general when it is difficult or impossible to make experimental measurements.

REFERENCES AND NOTES

(1) Ran, Y.; Jain, N.; Yalkowsky, S. H. Prediction of Aqueous Solubility of Organic Compounds by the General Solubility Equation (GSE). *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1208−1217.

(2) Abraham, M. H.; Le, J. The Correlation and Prediction of the Solubility of Compounds in Water Using an Amended Solvation Energy Relationship. *J. Pharm. Sci.* **1999**, *88*, 868−880.

(3) Yalkowsky, S. H.; Dannelfester, R. M. *The Arizona Database Of Aqueous Solubility*; College of Pharmacy, University of Arizona: Tucson, AZ, 1990.

(4) Morelock, M. M.; Choi, L. L.; Bell, G. L.; Wright, J. Estimation And Correlation Of Drug Water Solubility With Pharmacological Param-

eters Required For Biological Activity. *J. Pharm. Sci.* **1994**, *83*, 948−952.
(5) Prankerd, R. J.; McKeown, R. H. Physico-Chemical Properties Of Barbituric Acid Derivatives: IV. Solubilities Of 5,5-Disubstituted Barbituric Acids In Water. *Int. J. Pharm.* **1994**, *112*, 1−15.
(6) Williams, G. C.; Sinko, P. J. Oral Absorption Of The HIV Protease Inhibitors: A Current Update. *Adv. Drug. Deli. Rev.* **1999**, *39*, 211−238.
(7) Fichan, I.; Larroche, C.; Gros, J. B. Water Solubility, Vapor Pressure, And Activity Coefficients Of Terpenes And Terpenoids. *J. Chem. Eng. Data* **1999**, *44*, 56−62.
(8) Aungst, B. J. P-Glycoprotein, Secretory Transport, And Other Barriers To The Oral Delivery Of Anti-HIV Drugs. *Adv. Drug. Deli. Rev.* **1999**, *39*, 105−116.
(9) Kristl, A. Estimation Of Aqueous Solubility For Some Guanine Derivatives Using Partition Coefficient And Melting Temperature. *J. Pharm. Sci.* **1999**, *88*, 109−110.
(10) Myrdal, P. B.; Manka, A. M.; Yalkowsky, S. H. AQUAFAC 3: Aqueous Functional Group Activity Coefficients: Application To The Estimation Of Aqueous Solubility. *Chemosphere* **1995**, *30*, 1619−1637.
(11) Kristl, A.; Vesnaver, G. Thermodynamic Investigation Of The Effect Of Octanol−Water Mutual Miscibility On The Partitioning And Solubility Of Some Guanine Derivatives. *J. Chem. Soc., Faraday Trans.* **1995**, *91*, 995−998.
(12) Kristl, A. Thermodynamic Investigation Of The Effect Of The Mutual Miscibility Of Some Higher Alkanols And Water On The Partitioning And Solubility Of Some Guanine Derivatives. *J. Chem. Soc., Faraday Trans.* **1996**, *92*, 1721−1224.
(13) Stella, V. J.; Martodihardjo, S.; Rao, V. M. Aqueous Solubility And Dissolution Rate Does Not Adequately Predict In Vivo Performance: A Probe Utilizing Some N-Acyloxymethylphenytoin Prodrugs. *J. Pharm. Sci.* **1999**, *88*, 775−779.
(14) Cerius^2 (Combi-Chem, Compound Selection Procedures). Accelrys Inc., San Diego, CA, 2002, see http://www.accelrys.com.
(15) MOE. Chemical Computing Group Inc., Montreal, Quebec, Canada, 2002, see http: //www.chemcomp.com.
(16) Labute, P. A widely applicable set of descriptors. *J. Chem. Computing Group*.
(17) Wildman, S. A.; Crippen, G. M. Prediction of Physicochemical Parameters by Atomic Contribution. *J. Chem. Inf. Comput. Sci*. **1999**, *39* (5), 868−873.
(18) Gasteiger, J.; Marsili, M. Iterative Partial Equalization of Orbital Electronegativity − A Rapid Access to Atomic Charges *Tetrahedron* **1980**, *36*, 3219.
(19) Huuskonen, J. Estimation of Aqueous Solubility for a Diverse Set of Organic Compounds Based on Molecular Topology. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 773−777.
(20) Tetko, I. V.; Tanchuk, V. Yu.; Kasheva, T. N.; Villa Alessandro, E. P. Estimation of Aqueous Solubility of Chemical Compounds Using E-State Indices. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1488−1493.
(21) PLS-Toolbox. Eigenvector Research, Inc., Manson, WA, 2002, see http://www.eigenvector.com.
(22) STATISTICA, Neural Networks. StatSoft Inc., Tulsa, OK, U.S.A., see http://www.statsoft.com.
(23) Graham, G. W.; Sarker, M.; Dunn, W. J.; Scott, D. R. UNIPALS: Software for Principal Components Analysis and Partial Least Squares Regression. *Tetrahedron Comput. Methodology* **1989**, *2* (6), 377−396.
(24) de Jong, S. SIMPLS: an alternative approach to partial least squares regression *Chemom. Intell. Lab. Syst*. **1993**, *18*, 251−263.
(25) Wold, S.; Kettaneh-Wold; Skagerberg, B. Nonlinear PLS modeling. *Chemom. Intell. Lab. Syst.* **1989**, *7*, 53−65.
(26) Wold, S. Nonlinear PLS modeling II: Spline inner relation (SPL_PLS). *Chemom. Intell. Lab. Syst.* **1992**, *14*.
(27) Stone, M.; Brooks, R. J. Continuum Regression: Cross-validated Sequentially constructed Prediction embracing Ordinary Least Squares, Partial Least Squares and Principal Components Regression. *J. R. Statist. Soc. B* **1990**, *52*, 337−369.
(28) Wise, B. M.; Ricker, N. L. Identification of finite Impulse Response Models using Continuum Regression. *J. Chemometrics* **1993**, *7* (1), 1−14.
(29) Poggio, T.; Girosi, T. Networks for approximation and learning. *Proc. IEEE* **1990**, *78*, 1481−1497.
(30) Leonard, J.; Kramer, M. K. Improvement of the back-propagation algorithm for training neural networks. *Comput. Chem. Eng.* **1990**, *14*, 337−341.
(31) Bishop, C. *Neural Network for Pattern Recognition*; Oxford-University Press: 1995.
(32) Patterson, D. *Artificial Neural Networks*; Singapore-Prentice Hall: 1996.

CI049797U