JOURNAL OF CHEMICAL INFORMATION AND MODELING

ARTICLE

pubs.acs.org/jcim

Determining the Degree of Randomness of Descriptors in Linear Regression Equations with Respect to the Data Size

Michael C. Hutter*

Center for Bioinformatics, Campus Building E2.1, Saarland University, 66123 Saarbrücken, Germany

ABSTRACT: Linear regression equations suffer from the curse of dimensionality that leads to overfitting and accidental correlation, particularly for small data sets and when many variables are present. This can lead to cases where descriptors based on random numbers exhibit higher correlations than actual descriptors. In this study, it was therefore investigated how high the degree of accidental correlation of a single descriptor can be with respect to the number of observations. On the basis of computer simulations for data sizes ranging from 7 to 500 observations, a formula was derived that expresses the degree of randomness (in percent) of a chosen descriptor depending on its correlation coefficient and the size of the data set. This allows one to determine a cutoff for the correlation below which descriptors can be discarded due to a high risk of chance correlation. Doing so, the number of eligible variables for the regression analysis can be reduced substantially. Corresponding applications are reported for several QSAR data sets of various sizes.

INTRODUCTION

Despite their widespread use, quantitative structure−activity relationships have faced repeated criticism over the past three decades.[1−7] A common problem is overfitting due to the use of too many variables in the regression equation.[5] This is, however, obvious and can easily be avoided. Much more severe and difficult to handle is the theoretical possibility that one or more descriptors exhibit high correlations simply by chance. Topliss and co-workers investigated the probabilities of obtaining accidentally high correlations in multiple regression analysis in the 1970s.[1,2] They applied artificial descriptors obtained from a random number generator and focused on the correlation coefficients with respect to the number of observations (number of data points) and variables being used. Later, Rencher and Pun as well as Livingston and Salt performed more extensive simulations.[3,6,8] The recent work of Rücker et al. and by the group of Katritzky investigated the outcome comparing actual descriptors, which were derived from the molecular structure, and random descriptors for real data sets.[9,10]

All of these results clearly demonstrate that the lower the number of observations, and the larger the count of available descriptors, the higher the chance of an accidentally high correlation will be. Thus, even very high correlations (above 0.95) can be simply due to the large number of available variables, in particular for small data sets, e.g., less than 10 observations. Regarding the vast amount of available descriptors (>1000) that can be readily computed by computer programs such as DRAGON or CODESSA,[11,12] these considerations may render most QSAR approaches for data sets containing less than 50 compounds meaningless, simply due to chance correlation. Likewise, the question arises of whether actual descriptors (that are derived from the molecular structure) lead to higher correlations than those that are purely random numbers.[10]

In any case, a reasonable connection between observed quantity and used descriptors should be obvious, or at least desirable. Among other issues, Cronin and Schultz stressed the physicochemical interpretability of descriptors that give rise to a mechanistic basis in QSAR equations for pharmacological end points, such as specific toxicity.[4] The interpretability of QSAR models is furthermore always a trade-off between the choice of method (e.g., linear regression equations vs nonlinear support vector regression) and the obtainable accuracy (with the immanent danger of overfitting). Thus, the QSAR model should not be a *black box*.[13] Overtraining, which is known from neural networks, can lead to situations where the resulting model is only applicable to a specific data set, whereas it fails for other, similar data. This might also be the reason for the phenomenon termed *activity cliffs*.[14] In other cases, strikingly high correlation was found, but no meaningful connection between cause and effect is apparent. One of the most popular examples is Sies' report on the correlation between pairs of brooding storks and the number of newborn babies (r^2 = 0.99, 7 observations).[15] More recently, Johnson demonstrated more of such strange correlations, e.g., that between imported lemons and highway fatalities (r^2 = 0.97, 5 data points).[7] Further examples were collected by Doweyko, who also discussed the reasonable upper limit of the correlation coefficient for biological data.[16]

So far, the dependence of chance correlation from combinations of multiple descriptors in linear regression equations with respect to the number of available variables has been studied systematically.[1−3,6,8] The remaining question is, how high can the accidental correlation of one single descriptor be, irrespective of the size of the descriptors pool but subject to the number of observations? Or, aiming for practical applications, is there a criterion to dismiss a given descriptor because it would lead to an accidental correlation at a chosen threshold, e.g., Pearson's correlation

Received: August 30, 2011
Published: October 30, 2011

coefficient of 0.4? For this purpose, actual QSAR data sets as well as artificial data obtained from random numbers were used in this study to derive a corresponding formula.

COMPUTATIONAL METHODS

The descriptors used for the actual data sets comprise a total of 129 variables including quantum chemical quantities. These were obtained from semiempirical AM1 calculations using a modified version of the program package VAMP.[17,18] Compounds were energetically optimized to a gradient norm below 0.1 kcal mol^{-1} $Å^{-1}$ using the default eigenvector Following algorithm.[19] All other descriptors, e.g., those for logP, the topological polar surface area,[20] and several drug-likeness indices, were computed with in house PERL scripts. The full list of descriptors has been reported elsewhere in detail.[21]

Artificial descriptors were computed using the random number generator function *rand* of PERL, which was initialized using the *time* variable in each run. Preliminary studies showed that the obtained distribution of random numbers does not differ from that of white noise derived from atmospheric noise (data not shown). Therefore, this routine is sufficient for the intended purpose. Artificial data sets were generated for data sizes of 5, 7, 10, 15, 20, 25, 30, 40, 50, 75, 100, 125, 150, 200, 250, 300, and 500 *y* values. For each *y* value, 10 000 artificial descriptors based on random numbers were computed. For the QSAR data sets containing experimentally determined *y* values (see below), 1000 random variables were computed on top of the actual descriptors.

In contrast to random numbers, the activities (*y* values) of experimental data are usually not evenly dispersed over their respective range. Therefore, it should be less likely to achieve accidentally high correlation using random descriptors. To account for this uneven distribution of *y* values, the experimental activities were sampled in 100 bins covering their range. Random numbers were now assigned to reflect the same distribution as the experimental *y* values: While constructing each artificial descriptor, the obtained value from the sequence of random numbers was accepted if the associated bin was not already covered. Otherwise, the next random number was tested until all bins were done. This procedure was likewise applied to the artificial data sets where the *y* values also consisted of random numbers.

Descriptors that showed zero variance within a data set were discarded for that particular set to prevent singularities while deriving the corresponding regression equation. For the stepwise forward regression, the MS-Windows version of OpenStat2 was used.[22] Pearson's correlation coefficients between the *y* values and the random descriptors of the artificial data sets were computed within the same PERL program as used for processing the random numbers.

To obtain the nonlinear fitting equation for the relation between correlation coefficient, data size, and percent randomness, the corresponding functionality of the *R* program was used.[23]

DATA SETS

The following sets of compounds were investigated, listed in ascending order of data points, where *n* denotes the number of observations: binding of aromatic sulfonyl ester derivatives to papain (Ester, $n = 7$), Table IB in ref 24; nonpancreatic secretory phospholipase A_2 inhibitors (PLA2, $n = 12$);[25] mutagenicity of halogen-containing furanones in the *Salmonella typhimurium* TA100 assay (MX, $n = 24$);[26] antifiliarial activity of antimycin analogues, also known as the Selwood data set (Selwood, $n = 31$);[27] inhibitory growth concentration of 3-aryloxazolidin-2-one antibacterials against *Staphylococcus aureus* SFCO-1a strain (OXA, $n = 60$);[28] and binding affinity to the benzodiazepine receptor (BZR, $n = 163$).[29,30] The full range of descriptors derived from the 3D structure of the substances was computed for all of these data sets. Experimental data and structures were taken from the cited references. For the OXA and the BZR sets, the activities and corresponding SMILES were used from the Supporting Information given in ref 31.

Since experimentally determined activities are applied as their reciprocal logarithmic form in usual QSAR equations, the term "activity" is used hereafter implying this mathematical transformation to avoid any confusion referring to *linear* correlation.

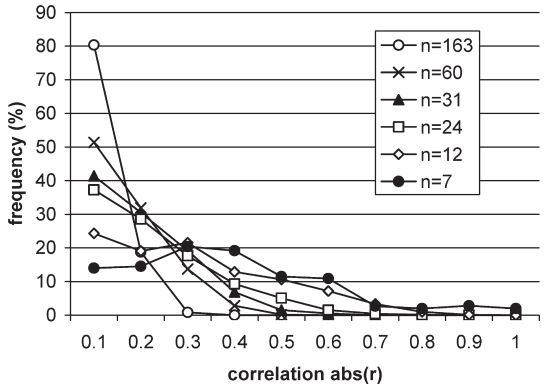

Figure 1. Probability distribution of the likelihood of obtaining an accidental correlation. Shown are the trends of the absolute correlation coefficients (Pearson) with the respective activity, computed from 1000 random descriptors in data sets comprising *n* observations.

RESULTS AND DISCUSSION

No algorithm for selecting variables while deriving regression equations can distinguish between actual descriptors, which have a connection to the observations, and purely random numbers. Furthermore, the descriptor that shows the highest correlation (r^2) with the activity will always appear in the regression equation irrespective of the algorithm being used. A random descriptor that exhibits a higher correlation than the best actual descriptor will therefore supersede the latter one. This is a truly unsatisfactory situation that should be avoided. For this purpose, we must know the probability that a random descriptor exhibits an accidentally high correlation (r^2) up to a given threshold for the particular data set. Therefore, simulations on six different data sets comprising 7−163 experimental observations were carried out to obtain trends, which were subsequently investigated more extensively on artificial data sets. Figure 1 shows that the chance of an accidentally high correlation (above 0.5) for one single descriptor is below 10% for data sizes larger than 10 observations. In general, the curves follow an exponential decay, except for the smallest two sets that show a maximum at 0.3. For $n = 7$, there is even a 5% chance of correlations of 0.9 and higher.

Considering only actual descriptors for this data set (Ester), SlogP[32] was identified as the variable being the most highly correlated with the experimental activity ($r = 0.987$), using it as the only descriptor in the regression equation ($r^2 = 0.974$, $F = 188.765$, se = 0.021). Obviously, the binding affinity is governed by the lipophilicity of the compounds as expressed by the computed water/*n*-octanol partitioning coefficient, if the uncertainty of an accidental correlation is accepted. Using the molar

dx.doi.org/10.1021/ci200403j |J. Chem. Inf. Model. 2011, 51, 3099–3104

refractivity (MR)[32] as a single variable, Hansch and Calef obtained r^2 = 0.501 and se = 0.392.[24] Here, MR showed a correlation of r = 0.735 with the activity, though being moderately intercorrelated to SlogP (r = 0.661).

For the phospholipase inhibitor set (PLA2, n = 12), the electrostatic hydrogen-bond acceptor capacity, termed ehbac,[33] was chosen as the only relevant descriptor in the stepwise forward procedure (r^2 = 0.536, F = 11.568, se = 0.731). The available X-ray structures show that binding is dominated by polar and ionic interactions, especially with the side chains of His48, Asp49, and Lys69 that are mediated by the charged oxygens of the acidic groups and the nitrogen atoms of the ligands.[25] Despite its moderate correlation with the binding affinity (r = 0.732), ehbac seemingly unites both electrostatic and hydrogen-bonding properties better than the remaining variables that were available. This is reflected by strong intercorrelation between ehbac and the count of nitrogen atoms (r = 0.920) and the sum of atomic charges on these atoms (r = −0.891).

MX compounds (MX, n = 24) are known to be strong electrophiles and therefore able to cause one-electron oxidation of guanine in the DNA in the absence of other reducing agents.[26] Thus, any variable that expresses electrophilic properties can be expected to correlate with the mutagenic activity, for example, the ionization potential (IP) and the energy of the lowest unoccupied molecular orbital (Elumo). Interestingly, the rugosity (rugos, defined as molecular volume devided by the surface area), as well as the covalent hydrogen-bond acidity (chbac)[33] are slightly more highly correlated with the activity (−0.888 and −0.878, respectively) than Elumo (−0.876) and therefore supersede the latter in the stepwise forward regression. The obtained statistical parameters for rugos alone are r^2 = 0.788, F = 81.222, and se = 1.660. Furthermore, chbac and Elumo are extremely highly intercorrelated (r = 1.000) and both strongly intercorrelated with rugos (0.900). Therefore, they are not selected as further descriptors in the stepwise forward approach. In the absence of rugos, chbac leads to slightly better values (r^2 = 0.771, F = 73.876, se = 1.728) than Elumo (r^2 = 0.768, F = 72.700, se = 1.738). Tuppurainen reported slightly different results for Elumo on the same set (r^2 = 0.865, F = 141.0, se = 1.33), which are possibly due to the use of a different optimization algorithm for the 3D structure of the compounds.[26]

Selwood and co-workers identified three descriptors as being relevant for the antifiliarial activity in their data set (Selwood, n = 31), namely melting point, logP, and the electrophilic superdelocalizability of certain atoms, in decreasing order of importance.[27] From the available descriptors in this study, again chbac showed the highest correlation with the activity (r = −0.690) and is likewise extremely intercorrelated with Elumo (r = 1.000) and strongly with SlogP (r = −0.816). Thus, chbac includes not only hydrogen-bond related terms but also hydrophobic information. Applying chbac alone, the obtained statistical parameters are r^2 = 0.476, F = 26.379, and se = 0.608, which are similar to Selwood's result for the training set (16 compounds) using only the melting point (r^2 = 0.49, F = 13.55, se = 0.58).

Karki and Kulkarni derived a series of QSAR equations for the interpretation of the antibacterial activity of the 60 compounds for which they collected experimental data from the literature (OXA, n = 60).[28] They found that Elumo appeared most often as a descriptor in the regression equations for the training set. Here, however, Elumo was not selected by the stepwise forward selection (r = −0.440). Instead, the Randić χ_0 index (chi0)[34] showed the highest correlation with the activity (r = 0.550) from all available variables for both the training set (50 compounds) and the whole set (n = 60), yielding r^2 = 0.302, F = 25.128, and se = 0.535 for the full set. The next two descriptors that were chosen were the drug-like index of Ghose et al. gvw80[35] (r = 0.434), indicating a substance to be within 80% of the preferred drugs, and the sum of atomic charges on nitrogen atoms, qsumn (r = 0.362). For the same training set, Katritzky and co-workers identified the average bond order for hydrogen atoms, the HOMO−LUMO energy gap, and the minimum electron−electron repulsion for the C−O bond as the most important descriptors.[36] The HOMO−LUMO energy gap denoted as dehl (r = −0.290) was also among the available variables here but is added much later to the regression equation. Obviously, the variability of the underlying variables in all of these approaches causes those descriptors to be included in the respective regression equation that shows the highest correlation with the activity in favor of similar but highly intercorrelated variables. Considering the size of this data set (n = 60), the chance of a random correlation of one single descriptor with the activity is below 3% for absolute values of r larger than 0.4 and below 0.2% if larger than 0.5, as can be seen in Figure 1. Thus, at least for the respective descriptor exhibiting the highest correlation, it can be safely assumed that this correlation is not by chance, regarding the number of observations.

For a data size of n = 163, as in the BZR set, the likelihood of a random correlation is even below 1% for absolute values of r larger than 0.3. Here, once again the covalent hydrogen-bond acidity (chbac) showed the highest correlation with the activity (r = −0.407). The obtained statistical parameters for the regression equation with this variable only are, however, poor (r^2 = 0.166, F = 32.049, and se = 1.009). Seemingly, this is an inhomogeneous data set, because the addition of further descriptors in the stepwise forward approach (number of protonatable NR_3 groups, Kier and Hall $^3\kappa_\alpha$ index,[37] and Kier and Hall $^6\chi$c connectivity index)[38] did not improve the correlation substantially (r^2 = 0.442, F = 31.304, and se = 0.833). For a subset of these benzodiazepines (55 compounds), it is known that electronegative substituents in positions 7 and 2′ increase the activity, whereas logP can strongly vary.[29]

Although it is apparent from the above real life examples that the likelihood of chance correlation decreases rapidly with growing data size, it would be interesting to see if there is a systematic dependency between the number of observations and the probability of accidental correlation of one single descriptor, which could be quantified by a mathematical formula. Such an expression would allow one to dismiss descriptors below a chosen level of confidence. Therefore, at first, systematic simulations on artificial data sets were performed, ranging from 5 to 500 y values, whereby 10 000 random descriptors were computed for each y value, as outlined in the Computational Methods section. These results are shown in Figure 2 for data sizes between 7 and 150. The shapes of the obtained curves are similar to those of the simulations for the actual data sets using 10 times fewer descriptor values (see Figure 1) and obviously follow a Gaussian distribution. To corroborate this assumption, a nonlinear fit was carried out for data sizes in the range of 10−125. For larger data sets, the curve decays even faster to zero, so that no meaningful data points for the fit are present. Conversely, data sizes below 10 show still large values even for correlations close to unity. Thus, the following fitting equation was obtained:

$$\text{randomness}\ (\%) = \frac{10\sqrt{2n}}{\sqrt{3}}\exp\left(\frac{-nr^2}{3}\right) \quad (1)$$

dx.doi.org/10.1021/ci200403j |J. Chem. Inf. Model. 2011, 51, 3099–3104

where n is the data size and *randomness* denotes the chance of having an accidental correlation of a descriptor that exhibits a squared correlation coefficient r toward the activity. To estimate how high a correlation coefficient must be, to be below a given threshold of accidental correlation, eq 1 can be rearranged:

$$r^2 = \frac{-3}{n} \ln\left(\frac{\text{randomness}\ (\%) \times \sqrt{3}}{10\sqrt{2n}}\right) \quad (2)$$

The following two examples should help in understanding the application of these formulas. First, let us consider how high the chance is that a descriptor showing a correlation $r = -0.505$ toward the activity is purely coincidental in a data set of size $n = 20$:

$$\text{randomness}\ (\%) = \frac{10\sqrt{2 \times 20}}{\sqrt{3}} \exp\left(\frac{-20(-0.505)^2}{3}\right) = 6.67\% \quad (3)$$

This means that the correlation of this descriptor is 100 − 6.67 = 93.33%, not at random, or in other words, at a confidence level of 0.1. Second, we want to know how large the correlation for descriptors with the activity must be in a data set containing 60 observation, if we request that the likelihood of accidental correlation is below 5% (which is equal to a confidence level of 0.05):

$$r^2 = \frac{-3}{60} \ln\left(\frac{5\sqrt{3}}{10\sqrt{2 \times 60}}\right) = 0.1269 \quad (4)$$

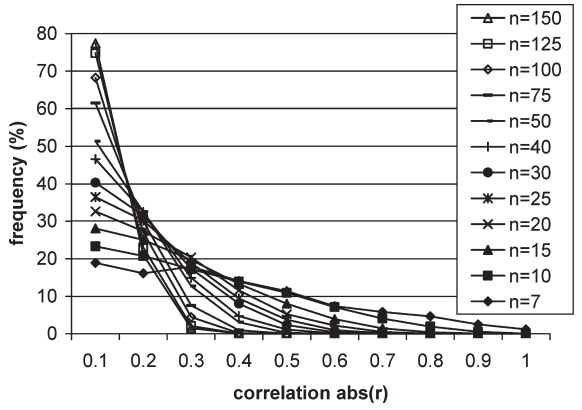

Figure 2. Probability distribution of the likelihood of obtaining an accidental correlation. Here, 10 000 random descriptors were computed for each y value in the artificial data sets of size n.

Accordingly, descriptors must exhibit squared correlation coefficients of $r^2 = 0.127$ or higher, which corresponds to values of $r = \pm 0.356$.

Descriptors can now be dismissed on the basis of eq 2 at a chosen percentage of being randomly correlated with the activity. This leads to a substantial reduction of the number of variables, depending on which degree of confidence is requested. Corresponding results for the six actual data sets are shown in Table 1. As expected, there is a monotonous decrease of the number of retained variables with an increasing level of confidence. For the two data sets containing the lowest number of observations (Ester and PLA2), no descriptors with less than 1% chance of accidental correlation remain. The trend for the MX data set is remarkable. It contains the lowest number of initially available descriptors, but the most at the highest level of confidence, with respect to its data size. A possible reason is that the compounds of the MX data set are rather homogeneous. The only nonhydrocarbon atoms appearing are oxygen, chlorine, and bromine. Therefore, many descriptors possess identical values for all of the molecules and are thus not considered due to zero variance right from the beginning. The opposite trend is present in the Selwood data set. Here, nearly all variables show variances unequal to zero initially, but only a few remain at higher levels of confidence. Regarding the size of this set, the chemical diversity seems to be much higher, comprising systematic variation of the substituents and their positions.

For all six data sets, the descriptor that is the most highly correlated with the activity is preserved at a confidence level of 0.05, for those sets of size $n > 12$ even at 0.01. Other descriptors, however, which are used as additional variables in the stepwise forward regression are dismissed due to insufficient correlation depending on the level of confidence. In the OXA set, for example, a confidence level of 0.01 (1% chance of accidental correlation) leads to a limit of $r^2 = 0.207$ ($r = \pm 0.456$). This causes gvw80 ($r = 0.434$, 1.5% chance correlation) and qsumn ($r = 0.362$, 4.6% chance correlation) to be dismissed, whereas chi0 ($r = 0.550$, 0.1% chance correlation) is still present. Most of the remaining variables above this margin exhibit high intercorrelations with chi0 that cause them to not be considered in the stepwise forward approach.

In this study, no attempt was made to divide the total data sets into training sets and test sets for the following reason: Since any subset of the total data set contains a smaller number of compounds, the chance of accidental correlation of the descriptors increases. Likewise, variables may exhibit high intercorrelations in subsets but not in the complete data set. This would cause those variables to be wrongly discarded due to incomplete information. Selection of appropriate descriptors according to

Table 1. Reduction of Descriptors in the Six Data Sets Using eq 2

		number of variables[a]	number of nonrandomly correlated variables at given percentages/levels of confidence			
data set	data size n	initially	80%, 0.20	90%, 0.10	95%, 0.05	99%, 0.01
Ester	7	101	86	54	41	0
PLA2	12	110	71	42	9	0
MX	24	93	69	46	36	22
Selwood	31	121	50	20	10	4
OXA	60	119	36	31	28	10
BZR	163	122	52	42	33	20

[a] For the count of initially available variables, those descriptors were removed that possessed zero variance.

dx.doi.org/10.1021/ci200403j |J. Chem. Inf. Model. 2011, 51, 3099–3104

eqs 1 and 2 should thus be done for the total data set and not for subsets.

In this context, it is certainly of interest how high the chance correlation of a corresponding regression equation comprising a given number of variables is. This probability, however, depends on the size of the descriptor pool, due to the chance of accidental correlation using a combination of two or more descriptors, as shown earlier.[1−3,6,8] Therefore, this situation has not been addressed. The focus of this study was instead to estimate the chance of accidental correlation for each descriptor on its own, which is independent from that of all other variables (c.f. the well-known formula for calculating Pearson's correlation coefficient). Thus, the derived eqs 1 and 2 are independent of the number of present variables. Moreover, they can be used to reduce the available descriptor pool substantially, and likewise the dimensionality of the best subset regression problem. Since the usual statistical tests to judge the significance of the resulting regression equations apply the F test, the corresponding values are also affected. In this context, Salt and Livingstone, along with other authors, pointed out the inappropriateness of the common F test and used Monte Carlo simulations instead to generate critical F values.[8] Earlier, Rencher and Pun derived corresponding values for r^2 from Monte Carlo simulations, which account for the upward bias during stepwise regression.[3]

A further problem seems to be the nature of the descriptors themselves. For the Sies data set ($n = 7$), where the number of brooding storks shows a correlation of $r^2 = 0.99$ with the number of newborn babies, the chance that this is an accidental correlation is thus only around 2% according to eq 1. Likewise, the mentioned correlation between imported lemons and highway fatalities ($n = 5$, $r^2 = 0.97$) is ca. 4% random. Therefore, either we have to believe in these relationships (as we would most likely do if these were quantitative structure−activity relationships involving molecules) or we have to obtain additional observations instead that would presumably render these variables as being much lesser correlated. These two examples show that is not trivial to give general recommendations of what cutoff separates random from actual correlations, particularly since both variables were computed to be relevant at a 95% level of confidence. Applying an even tighter margin would, however, cause substantially more descriptors to be discarded, as shown in Table 1. As a result, the number of remaining variables can become too small to set up a regression equation at all. Selecting a reasonable cutoff is thus a tradeoff between the statistical level of confidence and usefulness. Here, eqs 1 and 2 allow a fast estimate for the problem at hand.

CONCLUSIONS

The probability of chance correlation decreases fast with the number of observations in the data set. In small data sets comprising less than 10 observations, this can lead to a false conclusion about cause and effect despite overwhelming statistical evidence. For the derivation of structure−activity relationships, it is therefore advisible to use only reasonably large data sets. If the number of experimental data points is very large, it may, however, become difficult to find descriptors that exhibit high correlations at all. The derived formula to estimate the chance of accidental correlation can be used to reduce the descriptor pool substantially at a chosen level of confidence. This also reduces the computational complexity for any method upon performing regression analysis, for example, best subset selection.

AUTHOR INFORMATION

Corresponding Author

*Phone: +49 681 302 70703. Fax: +49 681 302 70702. E-mail: michael.hutter@bioinformatik.uni-saarland.de.

ABBREVIATIONS

F, Fisher value; se, standard deviation; r, Pearson's correlation coefficient

REFERENCES

(1) Topliss, J. G.; Costello, R. J. Chance Correlation in Structure-Activity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972**, *15*, 1066–1068.

(2) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Activity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.

(3) Rencher, A. C.; Pun, F. C. Inflation of R^2 in Best Subset Regression. *Technomet.* **1980**, *22*, 49–53.

(4) Cronin, M. T. D.; Schultz, T. W. Pitfalls in QSAR. *THEOCHEM* **2003**, *622*, 39–51.

(5) Hawkins, D. M. The Problem of Overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.

(6) Livingstone, D. J.; Salt, D. W. Judging the Significance of Multiple Linear Regression Models. *J. Med. Chem.* **2005**, *48*, 661–663.

(7) Johnson, S. R. The Trouble with QSAR (or How I Learned To Stop Worrying and Embrace Fallacy). *J. Chem. Inf. Model.* **2008**, *48*, 25–26.

(8) Salt, D. W.; Ajmani, S.; Critchton, R.; Livingstone, D. J. An Improved Approximation to the Estimation of the Critical *F* Values in Best Subset Regression. *J. Chem. Inf. Model.* **2007**, *47*, 143–149.

(9) Rücker, C.; Rücker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model.* **2007**, *47*, 2345–2357.

(10) Katritzky, A. R.; Dobchev, D. A.; Slavov, S.; Karelson, M. Legitimate Utilization of Large Descriptor Pools for QSPR/QSAR Models. *J. Chem. Inf. Model.* **2008**, *48*, 2207–2213.

(11) Dragon, version 6.0. http://www.talente.mi.it (accessed July 8, 2011).

(12) Katritzky, A. R.; Karleson, M.; Petrukin, R. *CODESSA PRO*; University of Florida: Gainesville, FL, 2005. http://www.codessa-pro.com (accessed July 7, 2011).

(13) Guha, R. On the Interpretation and Interpretability of Quantitative Structure-Activity Relationship Models. *J. Comput.-Aided Mol. Des.* **2008**, *22*, 857–871.

(14) Maggiora, G. M. On Outliers and Activity Cliffs-Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.

(15) Sies, H. A New Parameter For Sex Education. *Nature* **1988**, *332*, 495.

(16) Doweyko, A. M. QSAR: Dead or Alive? *J. Comput.-Aided Mol. Des.* **2008**, *22*, 81–89.

(17) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. The Development and Use of Quantum-Mechanical Molecular-Model: 76. AM1 - A new General-Purpose Quantum-Mechanical Molecular-Model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.

(18) Rauhut, G.; Alex, A.; Chandrasekhar, J.; Steinke, T.; Sauer, W.; Beck, B.; Hutter, M.; Gedeck, P.; Clark, T. *VAMP*; version 6.5; Oxford Molecular: Erlangen, Germany, 1997.

(19) Baker, J. An Algorithm for the Localization of Transition-States. *J. Comput. Chem.* **1986**, *7*, 385–395.

(20) Ertl, P.; Rohde, B.; Selzer, P. Fast Calculation of Molecular Surface Area as a Sum of Fragment Based Contributions and its Application to the Prediction of Drug Transport Properties. *J. Med. Chem.* **2000**, *43*, 3714–3717.

(21) Schneider, N.; Jäckels, C.; Andres, C.; Hutter, M. C. Gradual in Silico Filtering for Druglike Substances. *J. Chem. Inf. Model.* **2008**, *48*, 613–628.

(22) Miller, W. G. *OpenStat2*, version 6.1; W. G. Miller: West DeMoines, Iowa, 50265. http://www.statprograms4u.com/ (accessed Aug 29, 2011).
(23) *R*; version 2.13.0; R Development Core Team: Auckland, New Zealand, 2011. http://cran.r-project.org (accessed Apr 13, 2011).
(24) Hansch, C.; Calef, D. F. Structure-Activity Relationships in Papain-Ligand Interactions. *J. Org. Chem.* **1976**, *41*, 1240–1243.
(25) Schevitz, R. W.; Bach, N. J.; Carlson, D. G.; Chirgadze, N. Y.; Clawson, D. K.; Dillard, R. D.; Draheim, S. E.; Hartley, L. W.; Jones, N. D.; et al. Structure-Based Design of the First Potent and Selective Inhibitor of Human Non-Pancreatic Secretory Phospholipase A_2. *Nat. Struct. Biol.* **1995**, *2*, 458–465.
(26) Tuppurainen, K. Frontier Orbital Energies, Hydrophobicity and Steric Factors as Physical QSAR Descriptors of Molecular Mutagenicity. A Review with a Case Study: MX Compounds. *Chemosphere* **1999**, *38*, 3015–3030.
(27) Selwood, D. L.; Livingstone, D. J.; Comley, J. C. W.; O'Dowd, A. B.; Hudson, A. T.; Jackson, P.; Jandu, K. S.; Rose, V. S.; Stables, J. N. Structure-Activity Relationships of Antifilarial Antimycin Analogues: A Multivariate Pattern Recognition Study. *J. Med. Chem.* **1990**, *33*, 136–142.
(28) Karki, R. G.; Kulkarni, V. M. Three-Dimensional Quantitative Structure-Activity Relationship (3D-QSAR) of 3-Aryloxazolidin-2-one Antibacterials. *Bioorg. Med. Chem.* **2001**, *9*, 3153–3160.
(29) Haefely, W.; Kyburz, E.; Gerecke, M.; Mohler, H. Recent Advances in the Molecular Pharmacology of Benzodiazepine Receptors and in the Structure Activity Relationships of Their Agonists and Antagonist. *Adv. Drug. Res.* **1985**, *14*, 165–322.
(30) Sutherland, J. J.; O'Brien, L. A.; Weaver, D. F. A Comparison of Methods for Modelling Quantitative Structure-Activity Relationships. *J. Med. Chem.* **2004**, *47*, 5541–5554.
(31) Mittal, R. R.; McKinnon, R. A.; Sorich, M. J. Comparison Data Sets for Benchmarking QSAR Methodologies in Lead Optimization *J. Chem. Inf. Model* **2009**, *49*, 1810–1820.
(32) Viswanadhan, V. N.; Ghose, A. K.; Rebankar, G. R.; Robins, R. K. Atomic Physicochemical Parameters for Three Dimensional Structure Directed Quantitative Structure-Activity Relationships. 4. Additional Parameters for Hydrophobic and Dispersive Interactions and Their Application for an Automated Superposition of Certain Naturally Occuring Nucleoside Antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.
(33) Cronce, D. T.; Famini, G. R.; De Soto, J. A.; Wilson, L. Y. Using Theoretical Descriptors in Quantitative Structure-Property Relationships: Some Distribution Equilibria. *J. Chem. Soc., Perkin Trans. 2* **1998**, 1293–1301.
(34) Randić, M. On Characterization of Molecular Branching. *J. Am. Chem. Soc.* **1975**, *97*, 6609–6615.
(35) Ghose, A. K.; Viswanadhan, V. N.; Wendoloski, J. J. A Knowledge-Based Approach in Designing Combinatorial or Medicinal Chemistry Libraries for Drug Discovery. 1. A Qualitative and Quantitative Characterization of Known Drug Databases. *J. Comb. Chem.* **1999**, *1*, 55–68.
(36) Katritzky, A. R.; Fara, D. C.; Karelson, M. QSPR of 3-aryloxazolidin-2-one Antibacterials. *Biorg. Med. Chem.* **2004**, *12*, 3027–3035.
(37) Kier, L. B. Distinguishing Atom Differences in a Molecular Graph Shape Index. *Quant. Struct.-Act. Relat.* **1986**, *5*, 7–12.
(38) Kier, L. B.; Hall, L. H. The Nature of Structure-Activity Relationships and their Relation to Molecular Connectivity. *Eur. J. Med. Chem.* **1977**, *12*, 307–312.