Sharpening the Toolbox of Computational Chemistry: A New Approximation of Critical *F*-Values for Multiple Linear Regression

Christian Kramer,[†,‡] Christofer S. Tautermann,[‡] David J. Livingstone,[§] David W. Salt,[||] David C. Whitley,[§] Bernd Beck,[‡] and Timothy Clark*[,†,§]

Computer-Chemie-Centrum and Interdisciplinary Center for Molecular Materials, Friedrich-Alexander Universität Erlangen-Nürnberg, Nägelsbachstrasse 52, 91052 Erlangen, Germany, Department of Lead Discovery, Boehringer-Ingelheim Pharma GmbH & Co. KG, 88397 Biberach, Germany, Centre for Molecular Design, School of Pharmacy and Biomedical Sciences, University of Portsmouth, Mercantile House, Hampshire Terrace, Portsmouth PO1 2EG, U.K., and Department of Mathematics, Lion Gate Building, Lion Terrace, University of Portsmouth PO1 3HF, U.K.

Received September 3, 2008

Multiple linear regression is a major tool in computational chemistry. Although it has been used for more than 30 years, it has only recently been noted within the cheminformatics community that the standard *F*-values used to assess the significance of the resulting models are inappropriate in situations where the variables included in a model are chosen from a large pool of descriptors, due to an effect known in the statistical literature as selection bias. We have used Monte Carlo simulations to estimate the critical *F*-values for many combinations of sample size (n), model size (p), and descriptor pool size (k), using stepwise regression, one of the methods most commonly used to derive linear models from large sets of molecular descriptors. The values of n, p, and k represent cases appropriate to contemporary cheminformatics data sets. A formula for general n, p, and k values has been developed from the numerical estimates that approximates the critical stepwise *F*-values at 90%, 95%, and 99% significance levels. This approximation reproduces both the original simulated values and an interpolation test set (within the range of the training values) with an R^2 value greater than 0.995. For an extrapolation test set of cases outside the range of the training set, the approximation produced an R^2 above 0.93.

INTRODUCTION

Multiple linear regression (MLR)[1,2] is one of the most fundamental methods used in statistical model building. It fits the parameters of a linear equation to a data set, using mathematics that has long been known and is well established. A major benefit of MLR is that it provides linear equations that are easy to understand and interpret. Despite its age, MLR is still widely used in applications such as quantitative structure property relationships (QSPR), quantitative structure activity relationships (QSAR), and further *in silico* methods.[3] MLR is also used as a fitting procedure for force field parameters, scoring functions, etc.

The more parameters that are available for fitting, the better the fit is likely to be, but the risk of chance correlations with the response variable increases with the number of descriptors available.[4,5] In today's QSAR and QSPR modeling procedures, where hundreds or even thousands of descriptors are available,[6] descriptor selection plays a key role.[7−12] For MLRs in which all available descriptors are included, the statistical significance of a model is assessed in terms of the *F*-value, which is required to lie in the upper tail of the *F*-distribution.[13] In a review[14] of variable selection procedures in computational chemistry, Livingstone and Salt noted that the critical *F*-values that have been used for 30 years by computational chemists are not appropriate when the descriptors used in MLR equations have been selected from a much larger pool of descriptors.[15] The standard *F*-values do not take account of selection bias, an effect that inflates the *F*-values required for a particular significance level. Thus, the power and influence of chance descriptors has often been underestimated, in some cases substantially.

The distribution of *F*-values for the situation when the MLR equation contains only a subset of the available descriptors is referred to as F_{max} by Salt and Livingstone. No analytical derivation of F_{max} is currently known, but a large number of simulations were performed to produce estimates of the critical F_{max} values for a range of sample size (n), model size (p), and descriptor pool size (k). However, these simulations are based on exhaustive best subsets regression, which is not feasible when

$$ {}^{k}C_{p} = \frac{p!}{k!(p-k)!} \tag{1} $$

the number of ways of selecting k from p variables is large. Consequently, F_{max} values are unavailable for many data set sizes relevant to contemporary practice in computational chemistry.

When the size of the data set precludes best subset regression, a commonly used alternative is stepwise regression.[16] As this is a much faster procedure than all subsets

* Corresponding author fax: (+49)9131-852-6565; e-mail: tim.clark@chemie.uni-erlangen.de.
† Friedrich-Alexander Universität Erlangen-Nürnberg.
‡ Boehringer-Ingelheim Pharma GmbH & Co. KG.
§ School of Pharmacy and Biomedical Sciences, University of Portsmouth.
|| Department of Mathematics, University of Portsmouth.

10.1021/ci800318q CCC: $40.75
Published on Web 12/23/2008

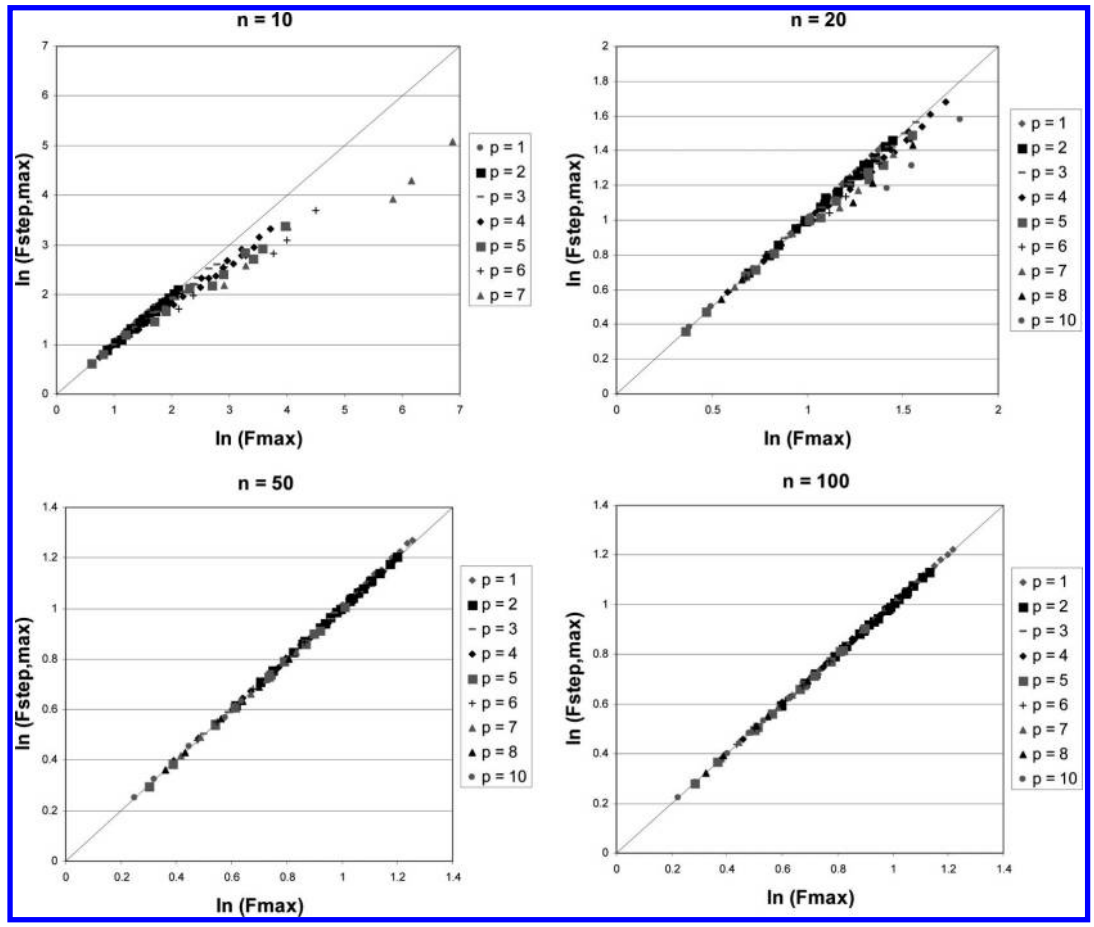

Figure 1. ln(F_{max})-values from best subset regression versus those from forward selection for $n = 10, 20, 50, 100$.

Table 1. Examples of $F_{max,step}$ -Values and Corresponding R^2 Values for Different n, k, and p

n	k	p	$F_{0.90}$	$R^2_{0.90}$	$F_{0.95}$	$R^2_{0.95}$	$F_{0.99}$	$R^2_{0.99}$
10	200	1	34.23	0.811	42.63	0.842	70.51	0.898
2000	5	1	5.33	0.00266	6.61	0.00330	9.73	0.00485
2000	200	1	12.06	0.00600	13.39	0.00666	16.43	0.00816
10000	500	1	13.72	0.00137	15.04	0.00150	18.29	0.00183
10	35	2	29.35	0.893	38.62	0.917	68.24	0.951
20	35	2	12.40	0.593	14.62	0.632	20.82	0.710
10	10	3	15.32	0.885	21.74	0.916	43.27	0.956
200	200	10	8.29	0.305	8.76	0.317	9.75	0.340
20	200	10	1.805	1.000	2.425	1.000	4.313	1.000
2000	3000	10	11.93	0.057	12.30	0.058	13.02	0.061

regression, the distribution of *F*-values for stepwise regression, denoted by $F_{max,step}$, can be simulated for a far wider range of n, p, and k. The resulting critical values are applicable directly to models derived using stepwise regression, and as best subsets regression always produces a model that fits at least as well as the model obtained by stepwise regression, the critical values of $F_{max,step}$ also provide lower bounds for the critical values of F_{max}. For small sample sizes some simulation data sets may contain models with reasonable R^2 values, and so F_{max} may deviate from $F_{max,step}$ as k and ${}^kC_p$ increase. However, for moderate and larger values of n, most simulations will produce poorly fitting models and $F_{max,step}$ is likely to be a good approximation to F_{max} and for $p = 1$,k $F_{max,step}$=F_{max}. We have therefore estimated the 90th, 95th, and 99th percentiles of the $F_{max,step}$ distribution for a range of n, p, and k that includes the values for which F_{max} estimates are available, so that the two may be compared, and extends beyond this to cover the sizes of typical QSAR/QSPR data sets. We also provide power series approximations of the critical $F_{max,step}$-values using seventeen terms. The terms are the same for all three significance levels; they are simple to integrate into programs running MLRs. For all $F_{max,step}$ estimates, the corresponding R^2 values are also provided, as these are perhaps more easily interpreted.

Previous Work. Rencher and Pun[17] and Diehr and Hoflin[18] provided approximation formulas for the maximum R^2 achievable with given combinations of n, k, and p. However, this work was carried out in the 1970s and 1980s, before simulated *F*-values for the ranges of n, k, and p commonly found in contemporary QSAR and QSPR models became available. Rencher and Pun for example simulated ranges of $n \leq 60$, $k \leq 40$, and $p \leq 10$ with stepwise regression.[17]

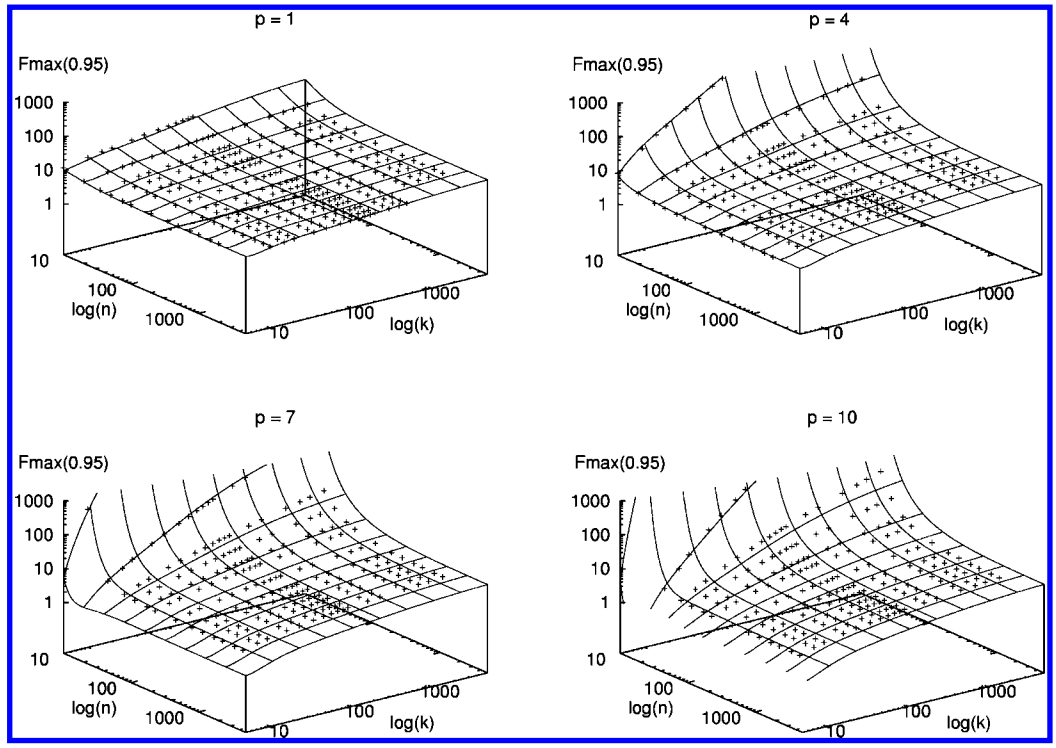

Figure 2. $F_{max,step}$-values and fitted surface for $p = 1$, 4, 7, and 10.

Table 2. Coefficients for the Approximation Formulas for All Three Significance Levels

element	coefficient	$P = 0.90$	$P = 0.95$	$P = 0.99$
$t_1 = 1$	α_1	3.16061	3.24277	3.40986
$t_2 = k^{-0.5}$	α_2	−12.1839	−11.6114	−10.5985
$t_3 = n^{-1}$	α_3	2.33692	3.30163	5.50809
$t_4 = p$	α_4	−0.116010	−0.132598	−0.166597
$t_5 = k^{-1}$	α_5	74.3387	72.4252	69.0419
$t_6 = p^2$	α_6	0.00732764	0.00834977	0.0104527
$t_7 = pk^{-0.5}$	α_7	−0.378087	−0.374248	−0.365010
$t_8 = pn^{-1}$	α_8	9.48873	9.63316	9.89630
$t_9 = k^{-1.5}$	α_9	−219.099	−213.703	−205.224
$t_{10} = pn^{-2}$	α_{10}	74.4535	75.5158	77.3272
$t_{11} = pn^{-1}k^{-0.5}$	α_{11}	−82.9063	−84.3390	−87.1630
$t_{12} = pn^{-1}k^{-1}$	α_{12}	275.807	282.228	295.279
$t_{13} = pn^{-3}$	α_{13}	−690.966	−702.111	−712.055
$t_{14} = k^{-2}$	α_{14}	222.304	216.939	209.328
$t_{15} = pn^{-1}k^{-1.5}$	α_{15}	−294.170	−303.400	−322.361
$t_{16} = p^2n^{-3}$	α_{16}	212.204	226.479	254.142
$t_{17} = p^2n^{-2}k^{-0.5}$	α_{17}	−50.6316	−51.7312	−53.6801

Livingstone and Salt[14] carried out extensive simulations of the F_{max} distribution using best subsets regression and fitted a formula for the 95th percentile of the F_{max} distribution to the simulated values. However, this formula gives poor estimates for values of n, p, and k outside the range of the simulations ($n \leq 100$, $k \leq 150$). Salt et al.[19] introduced an inflation index that describes the ratio of the critical F_{max}-values to those of the standard F-distribution. A complex expression for the inflation index was fitted empirically to simulated values of F_{max}. This approach is limited in practice by the time required to perform the all subsets regression; simulations for combinations of p and k where ${}^kC_p$ is large are infeasible.

Ruecker, Ruecker, and Meringer[20] showed in this journal how to use X- or Y-randomization to simulate the R^2 values that would be achievable by chance correlations. Assuming normal distribution of the R^2 values, they suggest that if the R^2 value achieved by the model is better than the average random R^2 plus n-fold standard deviation, it can be considered significant. In a courageous attempt, they judge 16 recently published QSAR and QSPR equations with respect to their significance.

Ruecker, Ruecker, and Meringer observed that the equations published for the F_{max}-value do not cover the range of n and k of contemporary QSAR/QSPR modeling and will not do so in the near future because of the very large number of models that must be calculated for higher n and k. While we agree on the first part, we now show that stepwise (rather than the more time-consuming best subset) MLR is a good approximation for higher n, and it is thus suitable for running simulations to extend the F_{max} formula to the ranges of n, p, and k common in current QSAR/QSPR modeling.

The Significance Problem. In multiple linear regression an equation of the type

$$y = a_1x_1 + a_2x_2 + ... + a_px_p + \varepsilon \quad (2)$$

where y is the response variable, $x_{1\text{-}p}$ are the descriptors, and ε is a random error is fitted to a data set by solving the normal equations. For a given set of y's and $x_{1\text{-}p}$'s, one optimal solution α_1-α_p is always found if $n > p$ and $x_j \neq \sum_{i \neq j} a_i x_i$.

The more descriptors available, the better the fit becomes. Even random descriptors may improve the fit. In order to judge whether a calculated solution to eq 2 is significant, it is compared to the result obtained using random descriptors. Depending on the choice of significance level, the result must be better than a predefined percentile of all results based on random descriptors. Commonly used values for the percentile are 90%, 95%, and 99%. The quality of the solution is compared to the result of a random solution in terms of R^2 or the F-value.

The coefficient of determination R^2 is defined as

$$R^2 = \frac{SS_{reg}}{SS_{tot}} = 1 - \frac{SS_{err}}{SS_{tot}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2} \quad (3)$$

Here SS_{reg} is the regressed sum of squares, SS_{tot} is the total sum of squares, SS_{err} is the sum of squared errors, y_i and $\hat{y}_i$ are the observed and predicted values of the i-th observation, and $\bar{y}$ is the mean value of all observations.

Definition of *F*-Values. For a solution of a regression task the F-value can be calculated as

$$F = \frac{\nu_2 R^2}{\nu_1 (1 - R^2)} \quad (4)$$

Here $\upsilon_1 = p$ and $\upsilon_2 = n\text{-}p\text{-}1$ are the first and second degrees of freedom. The F-value between two steps q and $q+1$ of a stepwise multiple linear regression can be calculated as

$$F = \frac{R^2_{q+1} - R^2_q}{(1 - R^2_{q+1})/(n - q - 2)} \quad (5)$$

Note that eq 5 reduces to eq 4 in the case of $q = 0$ and $p = 1$. F-values for given n and p can be found in common F-tables. However, the F-values tabulated do not consider the case where p descriptors are chosen from a pool of k descriptors, which increases the chance of finding better solutions. Thus, it is not possible to use the usual F-values in stepwise MLR and best subset MLR. Note that this situation applies to all variable selection algorithms.

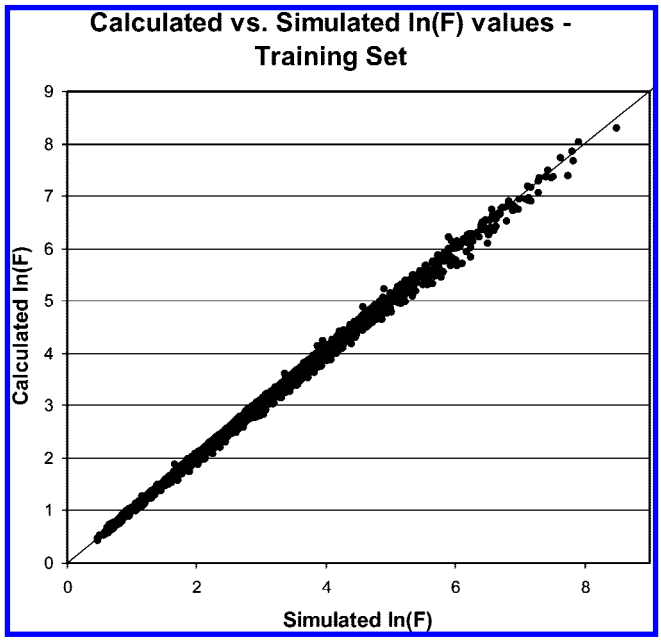

Figure 3. Plot of predicted versus simulated $F_{max,step}$-values.

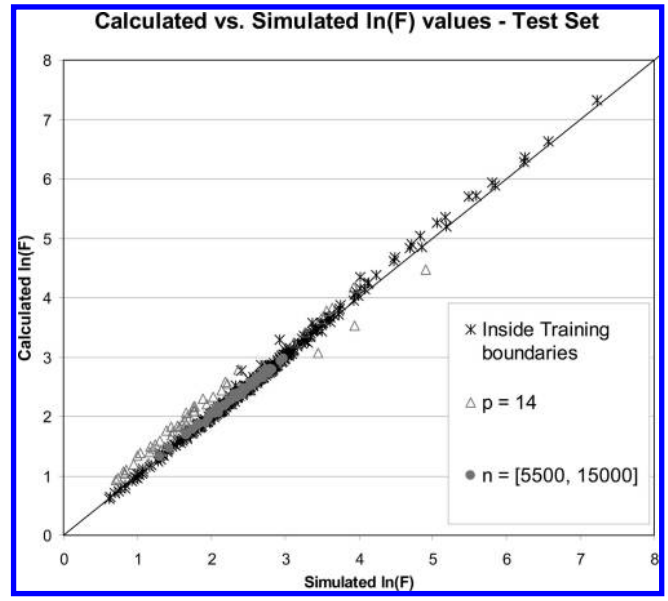

Figure 4. Plot of predicted versus simulated $F_{max,step}$-values for the test cases. Open triangles indicate values for $p = 14$.

Table 3. Coefficients for the Equation for Approximating $ln(F_{max,step})$ with $p = 1$

element	coefficient	$P = 0.90$	$P = 0.95$	$P = 0.99$
$y_1 = 1$	β_1	2.92371	2.99124	3.12901
$y_2 = k^{-0.5}$	β_2	−7.43273	−6.83193	−5.72839
$y_3 = n^{-1}$	β_3	12.3075	13.3999	15.9427
$y_4 = k^{-1}$	β_4	20.6095	19.3445	16.6906
$y_5 = n^{-1}k^{-0.5}$	β_5	−35.7882	−35.2874	−34.8941
$y_6 = k^{-1.5}$	β_6	−23.1200	−21.6866	−18.6847
$y_7 = n^{-1}k^{-1}$	β_7	42.6362	40.9349	39.3152

METHODS

Simulation of $F_{max,step}$-Values. The 90th, 95th, and 99th percentiles of the $F_{max,step}$ distribution were estimated for different combinations of the number of samples (n = 10, 20, 35, 50, 75, 100, 200, 350, 500, 750, 1000, 1500, 2000), the number of descriptors available (k=5, 10, 15, 20, 35, 50, 75, 100, 125, 150, 175, 200, 350, 500, 1000, 1500, 2000, 3000), and the number of descriptors chosen ($p = 1 - 10$). For each choice of (n,p,k) an array of n samples and k descriptors and a vector of n target values were filled with random numbers generated with the standard Fortran95 'random_number' function. A stepwise multiple linear regression up to p descriptors was then carried out using a forward-stepping algorithm.[21] Livingstone and Salt[19] showed that estimates of the critical F_{max}-values stabilize sufficiently after 50,000 repetitions and the same number of iterations was used here. Thus, the process was repeated 50,000 times for each (n,p,k) combination, the resulting R^2 and F-values were stored and sorted in ascending order, and the 45,000th, 47,500th, and 49,500th F-values were identified as estimates for the critical $F_{max,step}$ -values at 90%, 95%, and 99% significance levels.

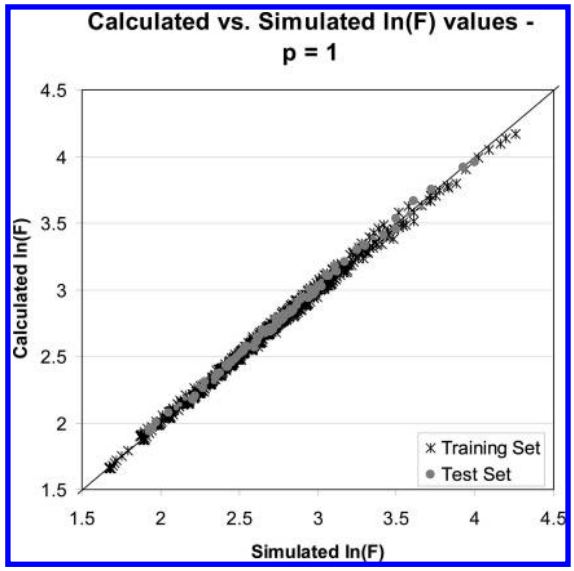

Figure 5. Plot of predicted versus simulated $F_{max,step}$-values for $p = 1$, training set black, test set gray.

The simulation error, i.e. the difference in $F_{max,step}$-values between two simulations for the same n, k, and p is $\leq 2\%$.

For higher n, the simulated $F_{max,step}$-values correlate surprisingly well with the simulated F_{max}-values from an exhaustive best subset search (taken from http://www.cmd-.port.ac.uk/cmd/fmax.shtml). We attribute this correlation to the linear independence and lack of correlation of random descriptors. A plot of $\ln(F_{max})$ versus $\ln(F_{max,step})$ is shown in Figure 1.

As expected, higher F_{max}-values for $n = 10$ tend to be underestimated. For $n = 20$, higher F_{max}-values also tend to be underestimated but not as much as for $n = 10$ and $p > 2$. The predictions for F_{max} values of $n = 50$ and $n = 100$ from best subset regression and from forward stepping MLR agree very well.

Approximation of $F_{max,step}$-Values. For a given number of descriptors p, the simulated $F_{max,step}$-values for different n and k appear to lie on a smooth hypersurface. The intention is to approximate the derived $F_{max,step}$ values by a global formula with variables n, k, and p employing a power series expansion of finite order. The asymptotic behavior of the simulated $F_{max,step}$ values, however, suggests first applying a variable transformation before expanding the power series. For large n and k values the $F_{max,step}$ values decrease monotonically toward a certain limit, implying that a transformation to negative exponents (e.g., $n \rightarrow n^{-1}$) might be a viable option. Therefore an expansion like

$$\ln(F_{max,step}(k,n,p)) = \sum_{0\leq r,s,t} c_{r,s,t} * (k^{e_1})^r * (n^{e_2})^s * (p^{e_3})^t \quad (6)$$

for the approximation of the $F_{max,step}$ values is envisioned, where $c_{r,s,t} \in \mathrm{IR}$ are the coefficients of the power series, $r,s,t \in \mathrm{IN}$ are the power exponents, and e_i are the exponents corresponding to the prior variable transformation. The expansion is truncated at the fifth order due to practical reasons as the observed approximation results are then seen to be sufficiently accurate. In preoptimization runs, where all 56 terms in the power series were included to fit the data, the exponents e_i were seen to give best results (low RMSE) when $e_1 \approx -0.5$, $e_2 \approx -1$, and $e_3 \approx 1$ (data not shown). Therefore, the power series was eventually expanded in the transformed variables n^{-1}, $k^{-0.5}$, and p leading to

$$F_{\max,step}(k,n,p) = \exp\left(\sum_{\substack{0\leq r,s,t\leq 5 \\ r+s+t\leq 5}} c_{r,s,t} * k^{-r/2} * n^{-s} * p^t\right) \quad (7)$$

In the next step, the number of included terms in the approximation, i.e. terms where $c_{r,s,t} \neq 0$, is reduced. This is necessary to eliminate linear dependence of the monomials in the series expansion, which would lead to numerical instabilities in the results. By a simulated annealing approach the number of terms with $c_{r,s,t} \neq 0$ is minimized with the aim to keep the RMSE of the fit close to the result with all 56 terms. The optimization suggests that 17 terms are sufficient to obtain reliable results (data not shown), and the 17 nonzero coefficients and terms were also determined by a simulated annealing approach. The final objective function is a combination of a low RMSE and a small linear dependency of the included terms.

We chose to fit $\ln(F_{max,step})$ instead of $F_{max,step}$ because there are many $F_{max,step}$ values below 50 and only a few very high values. If $F_{max,step}$ is used for fitting, the high values dominate the fitting procedure, and the approximation for the more important low values becomes worse. $F_{max,step}$ values above 1000 were removed from the data set, because even as $\ln(F_{max,step})$ they dominate and destroy the fit. $F_{max,step}$ values above 1000 all have corresponding R^2 values greater than 0.997, which is very probably outside the range commonly obtained in computational chemistry models.

RESULTS

$F_{max,step}$-values for each of the 90th, 95th, and 99th percentiles were calculated for 2243 combinations of n, k, and p. The R^2 values illustrate the correlation that must be reached in order to obtain a significant equation. It is revealing that, for example at the lower end, in 5% of all chance situations the R^2 for $n = 10$ samples and $p = 3$ descriptors selected out of $k = 10$ possible descriptors exceeds 0.90. Even when selecting from a pool of $k = 35$ descriptors, the $R^2_{0.95}$ for a 2-descriptor equation is approximately 0.92. Adding some samples here decreases the R^2 needed for significance by far, as the $R^2_{0.95}$ for $n = 20$ samples and 2 descriptors chosen out of 35 is roughly 0.63. At the higher end of n, the risk of chance correlations is very low. The $R^2_{0.95}$ for $n = 2000$, $k = 3000$, and $p = 10$ is 0.06. For $n = 10{,}000$ and k=500, a situation commonly observed for logP calculations or company in-house data sets, the $R^2_{0.95}$ for $p = 1$, direct comparison of correlation of single descriptors with the response variable, is 0.0015. Some more examples of R^2_p and $F_{max,step}$-values are given in Table 1.

The equation obtained for $ln(F_{max,step})$ with backward elimination of the expansion terms is shown in eq 8, where α_i are the coefficients and t_i are the expansion terms. The equation can be reduced to 17 terms without losing information. The terms in the equations are identical for all three significance values.

$$F_{max,step}(n,k,p) = \exp\left(\sum_{i=1}^{17} \alpha_i t_i\right) \quad (8)$$

Plots of the calculated surfaces and simulated $F_{max,step}$-values for $p = [1; 4; 7; 10]$ are plotted in Figure 2, which shows that the simulated $F_{max,step}$-values follow the values on a hypersurface. The gradient of the surface is small for $n > k$.

The $F_{max,step}$-values grow faster than exponentially when $k \gg n$.

The coefficients α_1-α_{17} and the elements t_1-t_{17} are given in Table 2.

This formula fits the data very well; it is not possible to remove any more terms without losing accuracy. The coefficients all follow the same tendency for increasing p, so that the approximation is smooth in coefficient space. Extremely high $F_{max,step}$-values destroyed the fit of our equation, so we omitted all $F_{max,step}$-values >1000. This should not be a problem, as $F_{max,step}$-values >1000 correspond to R^2-values >0.997. The maximum deviation between the calculated and simulated values of $\ln(F_{step})$ is 0.412 and the root-mean-square error (RMSE) of the fit is 0.049 (R^2=0.997), which is close to the simulation error. A plot of predicted versus simulated $F_{max,step}$-values is shown in Figure 3.

In order to test the equation, we simulated further $F_{max,step}$-values both inside and outside the boundaries of those used to generate the equation. Inside the boundaries n = [15, 40, 88, 250, 880], k = [12, 40, 88, 160, 250], and p = [1,2,4,7,10] were used. Outside the boundaries, p = 14 and n = [5,500; 15,000] were used. We did not extend k beyond 3000 because of the time these calculations take. The correlation of predicted versus simulated values inside the boundaries is $R^2(ln(F_{max,step}))$ = 0.995 and RMSE($ln(F_{max,step})$) = 0.058. For n = [5500, 15000] the fit has $R^2(ln(F_{max,step}))$ = 0.994 and RMSE($ln(F_{max,step})$) = 0.028. For p = 14 the fit has $R^2(ln(F_{max,step}))$ = 0.932 and RMSE($ln(F_{max,step})$) = 0.284. A plot of predicted versus simulated $F_{max,step}$-values at all 3 significance levels for all test cases is shown in Figure 4.

Most $F_{max,step}$-values for p = 14 (which is outside the fitting range) are predicted to be slightly too high, as can be seen from Figure 4. However, the trend is retained, and the values remain close to the correct ones. The other three outliers are for the three significance levels simulated for n = 15, k = 12, and p = 10. They correspond to R^2-values of 0.965, 0.979, and 0.993.

Only the $F_{max,step}$-values for p = 1 are relevant for stepwise linear regression approaches. We therefore simulated $F_{max,step}$-values for n up to 10,000 for p = 1 and fitted a sum of powers of n^{-1} and $k^{-0.5}$ to $ln(F_{max,step}(p = 1))$. The resulting seven term expression is shown in eq 9, where β_i are the coefficients, and y_i are the remaining expansion terms.

$$F_{max,step}(n, k, p = 1) = \exp\left(\sum_{j=1}^{7} \beta_j y_j\right) \qquad (9)$$

The equation fits all the simulated values with p = 1 very well with $R^2(ln(F_{max,step}))$ = 0.995 and RMSE$(ln(F_{max,step}))$ = 0.024. The coefficients for eq 9 for different significance levels are given in Table 3.

For p = 1, 49 test cases with n = [15; 40; 88; 250; 880; 5,500; 15000] and q= [12; 40; 88; 160; 250; 700; 1000] were simulated. The test set is predicted with $R^2(ln(F_{max,step}),test)$ = 0.996 and RMSE($ln(F_{max,step}),test$) = 0.024. A plot for simulated versus predicted values for p = 1 training and test cases are given in Figure 5.

The test values fit very nicely, even those with n = 15,000. This suggests that the equation found has the power to extrapolate.

DISCUSSION

We simulated F-values for different combinations of n, k, and p that should cover the space of contemporary QSAR/QSPR models. This was necessary because no such table for stepwise MLR is available in the literature. We have fitted a polynomial surface to the simulated values that has a very high correlation between simulated and predicted values for both training cases and test cases. Thus, we believe that the equations given here provide a basis for the simple and quick implementation of the correct F-values for stepwise linear regression. Moreover, apart from cases of small sample size, the simulated F-values obtained here using stepwise regression agree remarkably closely with earlier results[15,19] using best subsets regression. Thus we believe the critical $F_{max,step}$-values provide a good approximation to the critical F-ratios for general MLR models with sufficiently large sample sizes.

The equations fit to $F_{max,step}$-values below 1000 very well with an RMSE of 0.05, which is very close or equal to the error of simulation. The largest deviations between simulated and calculated values are found for high $F_{max,step}$-values, which are not relevant for drug discovery. As soon as $F_{max,step}$-values become higher than 100, the corresponding R^2 is above 0.95. If fits must be that good, a purely statistical view of the results is probably no longer sufficient, and noise in measurements and descriptors becomes very likely to govern the remaining error (at least in cheminformatics). Thus, the problem area is factor of 10 away from the area of interest.

In order to test our equations, we simulated $F_{max,step}$-values for n and p inside and outside the ranges of n, p, and k used to train the model. For p = 14 the predictions become slightly worse; for all the other predictions, even the series with n = 5500 and 15,000, which are outside the range of the training data, there is no real loss of observable quality. Thus, we believe that the equations are very reliable.

One of the basic assumptions for using the simulated $F_{max,step}$-values is that the descriptors are linearly independent. Of course, this assumption is not met in standard cheminformatics data sets. One way to identify the correct $F_{max,step}$-value for data sets containing multicollinearities is to reduce the dimensions by a technique such as UFS[22] or PCA.[23] The number of reduced descriptors must be used in place of k. Another way to obtain the correct $F_{max,step}$-value for a specific data set is to permute the response randomly, calculate multiple linear regressions 50,000 times, and use the $F_{max,step}$-value from the desired percentile, as suggested by Salt et al.[19] and slightly differently by Ruecker, Ruecker, and Meringer[20] (however, this approach is very time-consuming for large data sets.)

Besides the simplicity of MLR results, MLRs are popular because they use a measure of significance as the stopping criterion instead of cross-validation. Using cross-validation for adjusting parameters became very popular with the advent of neural networks in the nineties.[24] This technique is implemented in many algorithms. However, in order to be able to use cross-validation, the training data set must be reduced. This is undesirable if only a small number of data points is available. Additionally, a very strict and careful modeling strategy must be employed when using cross-validation. For example, descriptor selection based on R^2_{cv} should result in different descriptor sets for each cross-validation training set. If descriptor selection is performed

stepwise, there is a danger of inadvertent overtraining. For an overview of these and other disadvantages associated with R^2_{cv} and Q^2, see ref 25. Significance levels as the stopping criterion provide a simple approach to avoid overfitting if the correct F-value is used.

CONCLUSIONS

$F_{max,step}$-values for different combinations of n, k, and p that should cover the space of contemporary QSAR/QSPR models have been simulated. An approximation formula has been developed that fits the training and test $F_{max,step}$-values very well. Two different formulas, one for stepwise regressions and one for the complete formula, are given. These formulas are easy to implement in any MLR program and provide a basis for descriptor selection and model development alternative to R^2_{cv} approaches with all their associated problems.

ACKNOWLEDGMENT

C.K. thanks Boehringer-Ingelheim Pharma GmbH for financial support.

REFERENCES AND NOTES

(1) Legendre, A. M. *Appendix: On the method of least squares.* In new methods for determining the orbits of comets with a supplement that contains several improvements on those methods and their application to two 1805 comets.
(2) Gauss, C. F. *Theory of the motion of celestial bodies in a conical section of the sun's environment*; Perthes und Besser: Hamburg, Germany, 1809.
(3) Livingstone, D. J. The Characterization of Chemical Structures Using Molecular Properties - A Survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195–209.
(4) Topliss, J. G.; Costello, R. J. Chance Correlations in Structure-Avtivity Studies Using Multiple Regression Analysis. *J. Med. Chem.* **1972**, *15*, 1066–1068.
(5) Topliss, J. G.; Edwards, R. P. Chance Factors in Studies of Quantitative Structure-Acticity Relationships. *J. Med. Chem.* **1979**, *22*, 1238–1244.
(6) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(7) Votano, J. R.; Parham, M.; Hall, L. M.; Hall, L. H.; Kier, L. B.; Oloff, S.; Tropsha, A. QSAR modeling of human serum protein binding with several modeling techniques utilizing structure-information representation. *J. Med. Chem.* **2006**, *49*, 7169–81.
(8) Tropsha, A. Annual Reports in Computational Chemistry. In *Variable Selection QSAR Modeling*. Martin, Y., Ed.; Elsevier: Oxford, 2006; Chapter 4, Section 7, Model Validation; pp 113−126.
(9) Deng, W.; Breneman, C.; Embrechts, M. J. Predicting Protein-Ligand Binding Affinities Using Novel Geometrical Descriptors and Machine-Learning Methods. *J. Chem. Inf. Model.* **2004**, *44*, 699–703.
(10) Olah, M.; Bologa, C.; Oprea, T. I. An automated PLS search for biologically relevant QSAR descriptors. *J. Comput.-Aided Mol. Des.* **2004**, *18*, 437–449.
(11) Katritzky, A. R.; Stoyanova-Slavova, I. B.; Gobchev, D. A.; Karelson, M. QSPR Modeling of Flash Points: An Update. *J. Mol. Graphics Modell.* **2007**, *26*, 529–536.
(12) Kramer, C.; Beck, B.; Kriegl, J.; Clark, T. A compound model for hERG blockade. *ChemMedChem* **2008**, *2*, 254–265.
(13) Fisher, R. A. On a distribution yielding the error functions of several well known statistics. *Proc. Int. Cong. Math.* **1924**, *2*, 805–813.
(14) Livingstone, D. J.; Salt, D. W. Variable Selection - Spoilt for Choice. In *Reviews in Computational Chemistry*; Lipkowitz, K., Ed.; Wiley-VCH: Hoboken, NJ, 2005; pp 287−348.
(15) Livingstone, D. J.; Salt, D. W. Judging the Significance of Multiple Linear Regression Models. *J. Med. Chem.* **2005**, *48*, 661–663.
(16) Efroymson, M. A. Multiple regression analysis. In *Mathematical Methods for Digital Computers*; Ralston, A., Milf, H. S., Eds.; Wiley: NY, 1960; Vol. 1, pp 191−203.
(17) Rencher, A. C.; Pun, F. C. Inflation of R2 in Best Subset Regression. *Technometrics* **1980**, *22*, 49–53.
(18) Diehr, G.; Hoflin, D. R. Approximating the Distribution of the Sample R2 in Best Subset Regressions. *Technometrics* **1974**, *16*, 317–320.
(19) Salt, D. W.; Ajmani, S.; Crichton, 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.
(20) Ruecker, C.; Ruecker, G.; Meringer, M. y-Randomization and Its Variants in QSPR/QSAR. *J. Chem. Inf. Model* **2007**, *47*, 2345–2357.
(21) Kubinyi, H. QSAR in Drug Design. In *Handbook of Chemoinformatics*; Gasteiger, J., Ed.; Wiley-VCH: Weinheim, Germany, 2003; Vol. 4, pp 1532−1554.
(22) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised Forward Selection: A Method for Eliminating Redundant Variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160–1168.
(23) Pearson, K. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philos. Mag.* **1901**, *2*, 559–572.
(24) Zupan, J.; Gasteiger, J. *Neural Networks in Chemistry and Drug Design*, 2nd ed.; Wiley-VCH: Weinheim, Germany, 1999; ISBN 3-527-29778-2.
(25) Golbraikh, A.; Tropsha, A. Beware of Q2. *J. Mol. Graphics Modell.* **2002**, *20*, 269–276.

CI800318Q