The Probability of Chance Correlation Using Partial Least Squares (PLS)

Matthew Clark and Richard D. Cramer III*

Tripos Associates, Inc., 1699 South Hanley Road, St. Louis, MO 63144, U.S.A.

Abstract

The frequency of chance correlation using partial least squares (PLS) has been measured experimentally for variously dimensioned data, comprising either completely random numbers, random numbers containing a perfect correlation within, and CoMFA field descriptors. This frequency, much lower than that for stepwise multiple regression, is maximal for datasets in which the number of descriptors equals the number of compounds, and surprisingly decreases indefinitely as the number of descriptors becomes much greater than the number of compounds. However, perfect correlations involving descriptor subsets are not detected by PLS if the number of irrelevant descriptors is excessive. In CoMFA applications, the probability of chance correlation is usually negligible. For example with 21 compounds a crossvalidated r^2 value greater than 0.25 will occur by chance in less than 5% of trials.

Abbreviations and Symbols: PLS, Partial Least Squares analysis; MR, Multiple Regression; S-MR, Stepwise Multiple Regression; CoMFA, Comparative Molecular Field Analysis; r^2_{CV}, crossvalidated r-squared

Key words: Partial least squares, chance correlation, stepwise regression, CoMFA, cross validation

1 Introduction

Because many important sets of chemical observations cannot be adequately described by existing theory, researchers often seek semi-empirical models in which such important but uncontrolled observations (dependent variables) may be predicted as a linear function of more accessible (independent) variables. In the field of medicinal chemistry such models are called quantitative structure-activity relationships or QSAR. The reliability of the conventional tool for generating such models, multiple regression (MR), is well-established for situations where the candidate independent variables are much fewer in number than the sets of observations (e.g., Hansch analysis). However, in the very common situation in which the candidate independent variables are numerous, the MR technique can only consider small subsets of these independent variables, usually selected by a stepwise technique. When the candidate variables are interrelated, the subset models are very sensitive to small data value changes and are unreliable predictors. On the other hand, as the variables become less interrelated, the MR models resulting from stepwise selection are increasingly likely to be artifactual or chance correlations, phenomenologically meaningless and useless for prediction.

Commonly the significance of an MR model is estimated by calculating its F-ratio (ratio of "explained" to "unexplained" variance in the dependent variable, weighted by the number of parameters used to derive the model) and consulting a statistical table to find the probability that a correlation as good or better might arise by chance. The suspicion that this classical F-test greatly underestimates the true probability of chance correlation when stepwise selection precedes a final model led Topliss and collaborators [1, 2] to investigate this probability directly by simulation studies. They repeatedly filled an array with randomly generated numbers, then applied a standard stepwise MR algorithm to generate the best linear equation relating one column to some subset of the others. Of course, in this situation any "good" correlations must be artifactual. As feared, the frequency of such chance correlations was found to be much higher than might have been expected either intuitively, or as measured by the F-test. For example, within tables of random numbers comprising ten rows and ten independent variables, more than half the runs yielded an r^2 of at least 0.5. Another intuitive expectation – that increasing the number of candidate variables always increases the frequency of chance correlation – was confirmed.

Klopman and Kalos [3] observed that real sets of independent variables tend to be collinear and thus to contain fewer opportunities for chance correlations than do random numbers. To assess the probability of chance correlation while retaining the distribution properties of "real data", they suggested randomly interchanging the dependent variable values. Numerous interchanges would destroy the "true" structure-activity relationship, so that the vast majority of any correlations observed would be a chance arrangement among values which in fact are not related. A few examples using stepwise MR suggested that their CASE structural descriptors could be used in large number without the very high risk of chance correlation found by Topliss, if additional rules-of-thumb about F-ratios for adding new independent variables were postulated.

Partial least squares (PLS) is an important new technique [4] which seeks such linear models by repeated operations on all of the independent variables together, rather than by successively including individual variables. Functionally PLS is a superset of MR, in that PLS can produce the same results as MR wherever MR is applicable. But in contrast with MR, PLS allows any number of variables, in either the dependent or the independent block, and is most useful and stable when the variables

* To receive all correspondence

0931-8771/93/0102-0137$5.00+.25/0

within either block are intercorrelated. The PLS algorithm is described further in the Discussion.

The complexity of a PLS model is determined by crossvalidation [5] after each of the repeated operations. Crossvalidation, a computationally intensive but simple and direct technique for assessing the predictive utility of a model, also known as jackknifing or leave-1(n)-out, involves rederiving a model after omitting one or more randomly chosen observations, predicting the omitted dependent values from the model and the omitted independent values, and iterating until every observation has been omitted and predicted just once, accumulating the prediction errors. A crossvalidated r^2 value (r^2_{CV}) is computed from the equation:

$$r^2_{CV} = (SD - Press)/SD$$

where SD is the sum of the squared deviations of each dependent value from the mean of all dependent values, and Press is the sum of the squared deviations between the actual and predicted dependent values computed during the cross validation runs. Such a r^2_{CV} is quite analogous to the familiar classical r^2, but measures the predictive ability of a model rather than its data-fitting ability. A r^2_{CV} value of zero indicates that the model predictions have an average error no better than uselessly "predicting" any unknown dependent value to be the mean of the known values. A r^2_{CV} less than zero indicates that the average error of prediction is greater than the standard deviation of the original dependent values, a predictive performance which is worse than useless.

Intuitively, the crossvalidation technique integral to PLS would seem to reduce substantially the probability of chance correlation relative to MR. However the probability surely is not reduced to zero [6]. For example, if a second column of random numbers happens to exactly duplicate a first column, the perfect, though chance, correlation observed would also be confirmed by crossvalidation.

The chance correlation issue is of special importance because of the increasing usage of the Comparative Molecular Field Analysis (CoMFA) technique [7], in which PLS is used to relate hundreds or thousands of variables describing the steric and electrostatic fields surrounding a molecule to a property such as biological activity, for as few as ten compounds. Even though the CoMFA correlations in practice often have predictive utility, a level of concern about chance correlation remains, considering the results of Topliss. (A previous study of chance correlation in CoMFA [8] considered only the effect of randomizing the independent variables – the molecular fields – by perturbing the orientations of one or more compounds.)

We therefore decided to directly assess the probability of chance correlation using PLS with crossvalidation, by simulation studies similar to those of Topliss and coworkers, as a function of the dimensionality of the columns (variables, observations) and rows (cases). We also investigated the probability of diluting a true correlation, by adding many columns of random data to a small number of perfectly intercorrelated columns. Finally, because CoMFA variables are much more intercorrelated than are random numbers, we directly examined chance correlation within the three successful CoMFA studies [7, 9, 10] both by interchanging correct dependent variable values and also by generating random dependent variable values.

While these studies were being assembled for publication, we became aware of similar studies being undertaken by Wakeling and Morris [11]. Except for a few interpretative details, it is reassuring that these independent studies using a different implementation of PLS yield results identical to those described below.

2 Methods

The PLS algorithm, the one included in the QSAR module of the Sybyl software, had been validated by the reproduction of previously published PLS results and of MR results from other commercial software such as SAS (since the Sybyl version of PLS transforms the PLS solution back into the original measurement space, its output is directly comparable with MR). Its crossvalidation procedure is complete, that is, each crossvalidation run starts with the initial input data. All studies were carried out within Sybyl 5.3 or Sybyl 5.5 on an Iris GT-240 or ESV workstation, procedures being automated by means of the Sybyl-Programming Language (SPL).

Each of a series of data tables of different sizes was repeatedly filled with uniformly distributed random numbers in the range

Table 1. [a]Dimensions of tables containing random numbers only.

# of Rows	# of Independent Variables														
5	3[b]	5[b]	10˙	15	20	25	30	35	40	45	50	100	150	200[a]	300[a]
10	3[b]	5[b]	10[b]	15[b]	20	25	30	35	40	45	50	100	150	200[a]	300[a]
15	3	5	10[b]	15[b]	20[b]	25	30	35	40	45	50	100	150	200[a]	300[a]
20	3[b]	5[b]	10	15[b]	20[b]	25	30	35	40	45	50	100	150	200[a]	
25					20[a]	25[a]	30[a]	35[a]							
30						25[a]	30[a]	35[a]							
35							30[a]	35[a]	40[a]						

[a] Table refilled and analyzed 200 times. All others refilled and analyzed 1000 times.

[b] May be compared directly with results from stepwise MR [2].

0.0 to 1.0 [12]. The rows of a table correspond to "observations" or, in QSAR, "compounds"; the columns correspond to independent variables. One additional column of random numbers served as the dependent or target variables. The sizes of the various tables studied and the number of trials for each size are summarized in Table 1. Since it was noted for smaller tables that the highest frequencies of high r^2_{CV} values occur when the number of rows in the table equals the number of independent variables, the larger tables of 25, 30, and 35 rows were generated only with a similar number of columns. Because all values were drawn from a population of uniform mean and standard deviation, no scaling was applied to any of the numbers before analysis.

For each table, PLS analysis was carried out by performing full crossvalidation (leave-1-out) for each successive component, until all possible components had been extracted. The r^2_{CV} value recorded for an analysis was the highest observed with **any** number of components (which might be higher than the r^2_{CV} value usually reported by PLS, because conventionally PLS halts at the **first** maximum in r^2_{CV}). Then the table was refilled with a new set of random numbers and the PLS analysis repeated. This procedure was repeated 1000 times for each table with less than 150 columns and 200 times for larger tables, whilst recording the number of times the r^2_{CV} exceeded the 0.00, 0.05, 0.10, 0.25, 0.50, and 0.80 thresholds and the average and maximum of all positive r^2_{CV} values.

In order to study the effect of noise on PLS detection of perfect correlations, 10-row tables were generated in which each dependent variable value was defined to be the sum of one or more independent variable values. Thus the dependent variable was perfectly correlated with one or more independent variables combined, but uncorrelated (except by chance) with most of the independent variables considered by PLS. (However, note that these correlating independent variables are still generated randomly and thus are not themselves intercorrelated.) The table dimensions and number of repetitions are shown in Table 3. PLS analyses of each table were performed as described above.

For the direct assessment of chance correlation in CoMFA, three representative CoMFA data sets were selected, including 21, 37, and 74 compounds (rows). The biological activity values were scrambled by performing (2.5 × # compounds) interchanges of randomly selected pairs of rows. PLS was performed under "standard" CoMFA crossvalidation conditions (five crossvalidation groups, extraction of five components, omission of any independent variables whose variance was less than 2.0, and "CoMFA STD" scaling of independent variables so that the total variance of steric and electrostatic field values was equal). The r^2_{CV} value recorded was the highest value reached before any decline in r^2_{CV} value (the usual convention in PLS). The scrambling-then-PLS cycle was repeated 400 times for the 21-compound data set and 200 times for the other two, the distribution of r^2_{CV} values being recorded as in the previous studies. A second study of the 21-compound dataset used as "dependent variable" values random numbers uniformly distributed between 0.0 and 1.0, with the same PLS conditions, repeated 400 times.

3 Results

The results of the purely chance correlation experiments are summarized in Table 2. For each table dimension studied (# rows × # independent variables) are given: the percentage of times the r^2_{CV} exceeded the various thresholds; the average r^2_{CV}; and the maximum r^2_{CV}. There are three main conclusions from these experiments.

1. Despite the stochastic nature of the process which generated these data, Table 2 shows clearly (by scanning the sub-table corresponding to some number of rows) that the probability of obtaining a r^2_{CV} of practical interest (greater than 0.25, for example) tends to be highest **when the number of independent variables is equal to the number of columns.**
2. In particular, and counter to expectations based on experience with MR, Table 2 demonstrates that **the probability of chance correlation becomes negligible if the number of independent variables sufficiently exceeds the number of rows.** For example, with a table of ten rows analyzed by PLS, to accept a result as significant at the 95% level of confidence one would need (conservatively):
   - More than 150 independent variables if the observed r^2_{CV} was 0.25 or greater;
   - Either less than about 5 or more than 20 independent variables if the observed r^2_{CV} was 0.5 or greater;
   - Any number of independent variables if the observed r^2_{CV} is 0.8 or more.
3. Finally, as would be expected, the frequency of chance correlation for any given number of independent variables declines rapidly as the number of rows increases (for example, all values in the first row of Table 2, beginning with "5 3", are higher than the corresponding values in the first of the rows of Table 2 that begins with "10").

These three conclusions are perhaps most evident when the data are graphed. For example, Figure 1 is a three-dimensional graph showing the frequency of a r^2_{CV} greater than 0.25 as a

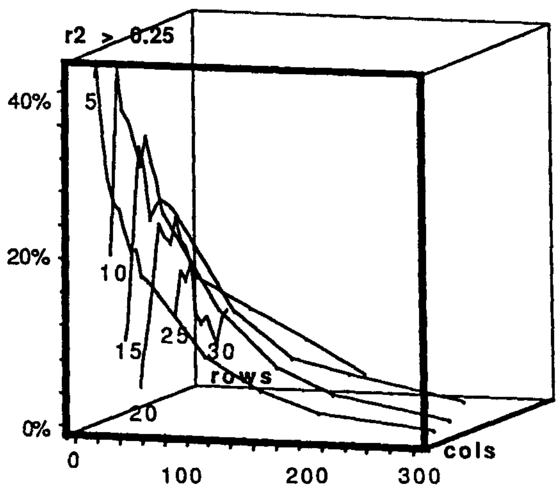

Figure 1. Frequency of a chance correlation with a r^2_{CV} value greater than 0.25, as a function of the numbers of rows and columns containing random data, using PLS. The figure is based on data in Table 1. See the text for further discussion.

Table 2. Results from PLS analysis of tables containing random numbers only

# Rows	# Indepdt Variables	% of r^2_{CV}:						Average Positive r^2_{CV}	Largest r^2_{CV}
		> 0.00	> 0.05	> 0.10	> 0.25	> 0.50	> 0.80		
5	3	28.3	27.0	25.8	20.6	15.0	7.2	0.150	1.000
5	5	28.3	27.1	26.3	21.1	12.3	4.1	0.132	0.992
5	10	25.9	23.7	22.2	16.9	8.0	1.8	0.098	0.942
5	15	24.7	22.2	19.7	14.4	6.1	0.3	0.081	0.850
5	20	25.1	22.7	19.7	12.8	4.6	0.5	0.072	0.920
5	25	26.9	23.7	20.2	12.5	4.8	0.3	0.076	0.898
5	30	24.9	22.8	19.5	11.3	3.0	0.0	0.066	0.745
5	35	25.4	22.4	18.4	10.1	2.6	0.2	0.061	0.895
5	40	23.7	21.6	18.0	10.2	2.8	0.1	0.061	0.882
5	45	23.0	19.8	17.0	8.7	1.3	0.1	0.052	0.882
5	50	24.0	20.8	16.7	8.4	1.7	0.0	0.051	0.795
5	100	19.8	15.2	10.7	4.0	0.2	0.0	0.029	0.571
5	150	14.3	10.7	7.3	2.1	0.1	0.0	0.019	0.522
5	200	10.0	6.6	3.6	0.9	0.0	0.0	.0.010	0.435
5	300	8.2	5.3	2.6	0.1	0.0	0.0	0.007	0.357
10	3	20.4	18.5	16.5	9.5	2.8	0.3	0.056	0.902
10	5	23.6	21.3	18.1	12.0	3.7	0.8	0.068	0.964
10	10	33.7	31.4	28.3	20.6	9.7	2.3	0.124	0.985
10	15	32.7	29.9	27.5	18.2	6.4	0.7	0.103	0.922
10	20	33.4	29.9	26.2	17.4	5.9	0.1	0.098	0.817
10	25	35.1	31.3	27.0	16.4	3.7	0.1	0.091	0.851
10	30	32.7	28.8	24.2	14.9	2.6	0.0	0.080	0.793
10	35	36.1	31.8	27.8	16.7	3.5	0.0	0.091	0.741
10	40	36.0	31.4	26.7	15.2	2.1	0.0	0.084	0.750
10	45	34.6	29.1	24.6	14.2	2.1	0.0	0.078	0.715
10	50	36.4	31.1	26.2	12.1	1.2	0.0	0.074	0.653
10	100	33.0	25.9	20.5	6.5	0.0	0.0	0.051	0.463
10	150	31.6	23.6	15.5	3.2	0.0	0.0	0.038	0.429
10	200	27.6	20.6	13.8	1.7	0.0	0.0	0.031	0.395
10	300	21.8	13.5	6.6	0.4	0.0	0.0	0.018	0.320
15	3	15.6	12.4	9.8	4.1	0.5	0.0	0.028	0.657
15	5	16.3	13.2	11.0	4.7	0.9	0.0	0.031	0.673
15	10	23.3	19.8	16.2	9.9	3.3	0.3	0.059	0.878
15	15	30.1	26.4	23.0	15.6	6.7	0.9	0.092	0.974
15	20	32.4	27.5	23.9	14.5	5.2	0.3	0.086	0.885
15	25	27.5	24.3	21.0	11.3	2.4	0.0	0.066	0.766
15	30	29.0	25.8	22.3	12.1	2.8	0.0	0.069	0.724
15	35	30.9	27.5	23.3	12.5	2.5	0.0	0.073	0.683
15	40	32.0	27.5	23.6	12.3	1.6	0.0	0.071	0.653
15	45	31.3	26.8	22.6	12.1	1.2	0.0	0.066	0.611
15	50	33.1	26.5	21.4	11.5	0.9	0.0	0.066	0.640
15	100	31.4	26.6	19.9	6.0	0.2	0.0	0.049	0.586
15	150	29.6	21.2	15.9	3.4	0.0	0.0	0.036	0.449
15	200	30.2	22.6	14.9	2.5	0.0	0.0	0.034	0.383
15	300	31.4	20.0	11.2	1.1	0.0	0.0	0.027	0.321
20	3	12.4	9.0	6.6	1.0	0.0	0.0	0.015	0.476
20	5	13.3	9.9	7.3	2.3	0.1	0.0	0.019	0.597
20	10	17.8	14.2	10.8	4.4	0.5	0.0	0.031	0.776
20	15	21.4	17.6	13.7	7.3	2.5	0.3	0.047	0.875
20	20	24.3	20.1	17.6	10.6	3.5	0.5	0.063	0.996
20	25	25.5	22.0	18.8	9.9	2.4	0.2	0.060	0.888
20	30	26.8	22.4	18.1	9.5	1.8	0.0	0.056	0.782
20	35	26.5	23.2	20.3	11.1	1.6	0.0	0.060	0.742
20	40	28.0	23.5	19.6	9.6	1.6	0.0	0.059	0.781
20	45	26.8	23.3	20.2	9.3	1.6	0.0	0.057	0.647
20	50	25.0	20.7	16.7	7.7	0.8	0.0	0.048	0.594
20	100	29.8	23.6	18.3	5.9	0.0	0.0	0.045	0.469
20	150	30.7	23.4	16.3	4.1	0.1	0.0	0.040	0.534
20	200	30.3	21.9	14.4	2.2	0.0	0.0	0.034	0.394
25	20	18.5	14.5	10.5	5.0	1.0	0.0	0.032	0.625

Table 2 continued. Results from PLS analysis of tables containing random numbers only

# Rows	# Indepdt Variables	% of r^2_{CV}:						Average Positive	Largest
		> 0.00	> 0.05	> 0.10	> 0.25	> 0.50	> 0.80	r^2_{CV}	r^2_{CV}
25	25	17.0	15.5	15.0	7.5	1.5	0.0	0.041	0.649
25	30	23.5	19.0	11.5	7.0	0.0	0.0	0.039	0.444
25	35	22.0	18.0	14.0	8.0	1.0	0.0	0.044	0.536
30	25	20.0	14.0	10.0	4.5	1.5	0.0	0.033	0.649
30	30	20.5	14.5	11.0	4.0	1.0	0.5	0.033	0.868
30	35	20.0	14.5	11.0	4.5	0.5	0.0	0.032	0.550
35	30	15.5	10.0	7.0	2.5	0.5	0.0	0.022	0.684
35	35	20.5	15.5	11.5	4.0	1.0	0.0	0.032	0.711
35	40	17.5	11.0	8.0	4.5	0.5	0.0	0.025	0.509

function of the number of rows and number of independent variables in the data table. The lines connect results for tables having equal numbers of rows. Particularly striking are the presence of a maximum in each of the lines and the steady decline in the frequency of chance correlation as the number of independent variables becomes very large. The decline in chance correlation as the number of rows increases is most easily appreciated by comparing the heights of the left **endpoints** of each of the lines (the apparent jump between 20 and 25 rows is because the smallest number of independent variables studied here changed from 3 to 20).

A direct comparison of the risk of chance correlation between PLS and stepwise MR is also desirable. There were thirteen table dimensions common to both our PLS study and the stepwise MR study by Topliss and Edwards [2], each noted with a (^{b}) in Table 1. Figures 2 and 3 plot the risk of chance correlation from stepwise MR against that for PLS over the thirteen common table sizes as two relevant indices, the mean classical r^2 and the frequency of observing a classical r^2 greater than 0.5. In these plots, the desirable region of low risk values is that closest to the origin. (Notice that the PLS and stepwise-MR axes are scaled differently, so that "equi-risk" lines have slopes of around 10 degrees rather than 45 degrees). The arrows connect tables having the same number of rows and point toward increasing numbers of independent variables.

Figures 2 and 3 show that, for all comparable table dimensions, the risk of chance correlation is much greater with stepwise MR than with PLS. For example, with 15 rows and 20 independent variables (the point at the tip of the arrow labelled 15), Figure 3 shows that the frequency of obtaining a classical r^2 greater than 0.5 by chance with stepwise MR is 75% (y coordinate of that point), about fifteen times as often as the frequency of 5% for r^2_{CV} with PLS (x coordinate of the same point).

The results of the experiments in which the dependent variable values equalled the sum of one to three independent variables are shown in Table 3. Because these tables intentionally contain a perfect correlation, the important question here is "can the noise introduced by some larger number of unrelated independent variables prevent PLS from recognizing a 'true' correlation present?" Table 3 shows that this risk certainly cannot be ignored. For example, the second row of the table shows that if there is one perfectly correlating column amid nine unrelated columns, PLS will report a correlation with a r^2_{CV} of 0.80 or

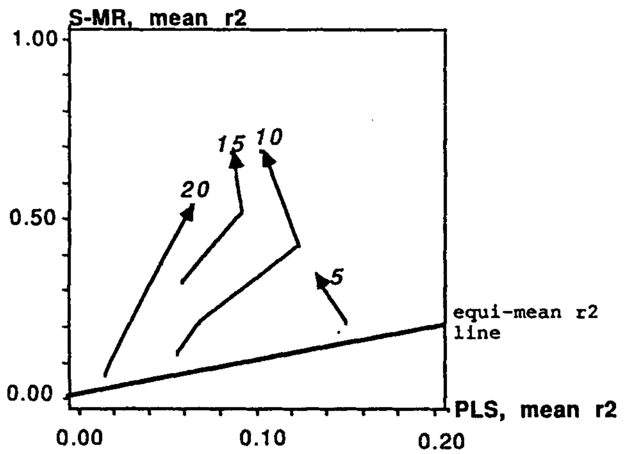

Figure 2. Comparison of the mean r^2 value derived by stepwise multiple regression with the mean r^2_{CV} value from PLS, for selected combinations of rows and columns containing random data. The figure is based on data from Reference 2 and Table 1. See the text for further discussion.

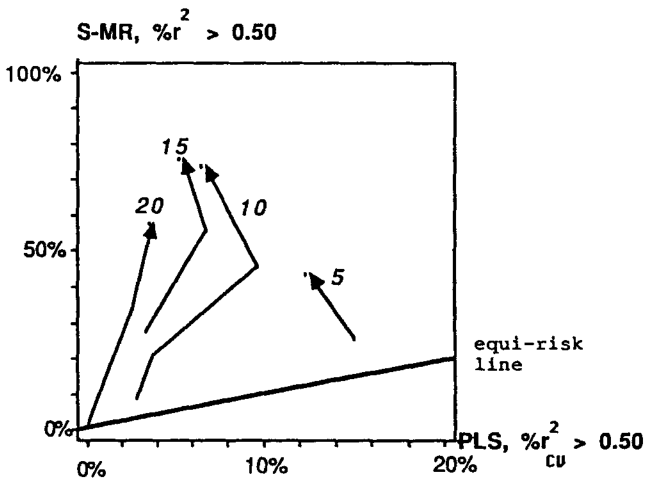

Figure 3. Comparison of the frequency of r^2 values greater than 0.5 derived by stepwise multiple regression with the frequency of r^2_{CV} values greater than 0.5 from PLS, for selected combinations of rows and columns containing random data. The figure is based on data from Reference 2 and Table 1. See the text for further discussion.

Table 3. PLS results from diluting tables containing a perfect correlation with random values

# Cols, sum Correlates	# Indepdt Variables	% of r^2_{CV}:						Average Positive	Largest
		> 0.00	> 0.05	> 0.10	> 0.25	> 0.50	> 0.80	r^2_{CV}	r^2_{CV}
1	5	100.0	100.0	100.0	100.0	100.0	100.0	1.000	1.000
1	10	97.7	97.5	96.9	94.9	87.3	55.4	0.764	0.998
1	15	90.6	88.7	86.6	78.9	56.7	10.1	0.491	0.969
1	20	83.4	80.7	77.7	64.2	36.2	2.8	0.367	0.924
1	25	79.1	74.9	70.7	57.7	25.4	1.7	0.312	0.882
1	30	73.9	70.4	66.6	51.4	18.6	0.2	0.269	0.847
1	35	70.7	66.7	61.5	43.4	13.4	0.1	0.229	0.840
1	40	68.4	64.6	59.5	43.0	12.7	0.1	0.221	0.844
1	45	66.1	61.1	55.5	35.6	8.9	0.1	0.191	0.857
1	50	64.6	58.5	52.9	33.8	7.3	0.0	0.178	0.730
1	100	52.7	44.9	35.3	15.0	0.5	0.0	0.096	0.576
1	150	46.0	36.1	27.3	7.8	0.1	0.0	0.065	0.536
1	200[a]	40.3	29.2	18.9	2.7	0.0	0.0	0.046	0.403
1	300[a]	35.6	24.9	14.3	0.9	0.0	0.0	0.034	0.399
2	3	100.0	100.0	100.0	100.0	100.0	100.0	1.000	1.000
2	5	100.0	100.0	100.0	100.0	100.0	100.0	1.000	1.000
2	10	97.7	97.0	96.4	93.0	86.0	55.1	0.745	1.000
2	15	89.5	87.9	85.1	76.1	52.1	10.9	0.474	0.984
2	20	81.7	78.6	75.1	62.4	35.1	2.4	0.358	0.925
2	25	77.5	74.1	70.8	55.9	27.0	0.9	0.309	0.906
2	30	69.1	66.6	61.5	45.2	16.8	0.3	0.245	0.883
2	35	68.8	64.7	60.4	43.5	12.1	0.0	0.226	0.757
2	40	66.8	63.0	58.5	40.3	12.3	0.0	0.213	0.785
2	45	66.4	61.8	55.3	38.8	9.8	0.0	0.198	0.774
2	50	62.4	57.6	50.7	31.0	7.1	0.0	0.168	0.786
2	100	52.5	45.0	37.7	15.4	0.9	0.0	0.097	0.593
2	150	44.5	33.8	25.6	6.3	0.1	0.0	0.061	0.580
2	200[a]	36.8	28.8	19.6	2.6	0.0	0.0	0.045	0.415
2	300[a]	33.2	22.7	14.2	1.2	0.0	0.0	0.032	0.369
3	5	99.9	99.9	99.9	99.9	99.9	99.9	0.999	1.000
3	10	97.3	97.0	96.1	94.0	85.3	53.3	0.743	1.000
3	15	87.7	85.8	83.5	74.0	49.8	11.8	0.461	0.974
3	20	81.8	79.5	76.0	63.8	37.1	3.2	0.371	0.906
3	25	75.0	71.5	67.5	54.6	23.4	0.8	0.293	0.872
3	30	70.2	67.3	62.8	46.5	18.3	0.2	0.250	0.817
3	35	69.9	65.0	59.6	41.6	14.4	0.1	0.223	0.857
3	40	68.6	63.2	56.4	39.8	9.3	0.0	0.201	0.780
3	45	65.9	60.7	53.5	36.5	8.7	0.1	0.190	0.813
3	50	61.9	56.2	50.3	31.1	4.7	0.0	0.159	0.712
3	100[a]	50.9	44.2	35.4	15.6	0.7	0.0	0.094	0.668
3	150[a]	43.2	34.8	28.0	7.0	0.1	0.0	0.066	0.606
3	200[a]	41.9	31.2	22.2	4.2	0.0	0.0	0.052	0.429
3	300[a]	38.5	25.1	15.1	1.2	0.0	0.0	0.035	0.317

[a] Experiment repeated 200 times. All other experiments repeated 1000 times.

better barely half of the time (55.4%). Even if one accepted correlations with a r^2_{CV} of only 0.25 or better, the fifth row shows that one perfectly correlating variable amid twenty would be overlooked by PLS almost half the time. In contrast, stepwise MR would find all of the perfect correlations involving a single independent variable and almost all of those involving two or three independent variables. A second important and perhaps unexpected result in Table 3 is that the probability of PLS recognizing a 'true' correlation is not affected by the number of variables needed to completely express that correlation (for example, all values within the same columns of the rows labelled 1-10, 2-10, and 3-10 within Table 3 are essentially identical).

The results of chance correlation studies with CoMFA appear in Table 4. Independent variables based on molecular fields are expected to be very much more collinear than random numbers, and so the central question is "Are the extensive results in Table 2 for random numbers at all applicable to CoMFA variables, or indeed to any sets of collinear variables?" The first three rows of Table 4 describe experiments in which correct dependent variable values were randomly scrambled, and the last row an experiment identical to the first row except the dependent variable values simply were random numbers. Comparison of the first and last rows suggests that the method of generating random dependent variables does not affect the frequency of chance correlation. Comparison of the first three

Table 4. CoMFA results with randomized dependent variables

Data Set Ref #	# Rows	# Ind Var	# Tries	True r^2_{cv}	% of r^2_{cv}:					Largest r^2_{cv}
					> 0.00	> 0.05	> 0.10	> 0.25	> 0.50	
[7]	21	105	400[a]	0.696	20.0	14.3	10.3	4.5	0.3	0.677
[9]	37	145	200[a]	0.417	15.0	8.5	3.0	0.0	0.0	0.215
[10]	74	156	200[a]	0.678	9.5	0.5	0.0	0.0	0.0	0.068
[7]	21	105	400[b]	0.696	17.3	14.0	10.0	3.0	0.5	0.517

[a] The "biological data" were randomly interchanged actual values.
[b] The "biological data" were random values.

lines shows, as expected, that more rows strongly reduces the frequency of chance correlation. To address the central question about the relevance of random number simulations to CoMFA, first notice that the maximum number of independent variables in crossvalidated CoMFA-based PLS is "only" about a hundred (column 3 of Table 4), rather than thousands, because around 90% of the lattice intersections have energy variances less than 2.0 and are omitted. Thus either the first or last line in Table 4 may be compared with the line in Table 2 beginning with "20 100". This comparison suggests that, if there is any difference between uncorrelated random numbers and correlated CoMFA field values in the frequency of chance correlations, chance correlation is more likely with random numbers. Are the CoMFA fields really much more intercorrelated than random numbers? One very satisfactory method for assessing intercorrelation of multivariate data is principal components analysis (PCA) (correlation coefficients, though often used for this purpose, measure only bivariate correlations and will not reveal substantial multivariate intercorrelations that may be present within multivariate data [13].) As an alternative to the determinant, and in the spirit of the F-ratio, we suggest as a "collinearity index", the ratio of "variance explained by the eigenvalues" to "variance unexplained by the eigenvalues", when some standard proportion of the possible eigenvalues, say half, has been extracted. The suggested variance ratio is easily calculated as "cumulative % variance explained"/(100-"cumulative % variance explained"). For example, from the independent variable values in the 21-compound CoMFA analyses cited in Table 4, twenty eigenvalues can be extracted. The cumulative % variance for the first 10 eigenvalues was 96.5%; the "collinearity index" for these data is 96.5/(100–96.5) = 27.6. For a conformant matrix of random numbers, the cumulative % variance after 10 eigenvalues was 66.9%; its collinearity index is 66.9/(100–66.9) = 2.0. In a very highly collinear dataset (see reference 13) the collinearity index is 165.7.

4 Discussion

These empirical findings raise several interrelated questions for discussion:

1) Why does PLS produce a relationship between number of independent variables and the probability of chance correlation which is bimodal, with a maximum when the number of independent variables equals the number of rows?
2) More columns of randomly generated independent variables certainly mean more chances for correlation with some dependent column. Why does PLS instead report fewer and fewer correlations as more independent variables than rows are added to a table?
3) What are the practical implications of these findings for structure-activity correlation work, using either stepwise MR or PLS, and in particular for CoMFA analyses?

The answers to the first two questions are intimately related to the behavior of the PLS algorithm. Qualitatively, PLS may be described as similar to principal components analysis (PCA). In both PCA and PLS, the fundamental operation is the repeated extraction of components, i.e., a linear combination of **all** variables under consideration, with each new component being orthogonal to any components previously, abstracted. But in PCA each new component includes only as much of the variance remaining in the variables as possible, whereas in PLS components are extracted, from both dependent and independent variable sets at the same time, in a way which expresses as much as possible of the variance which is common to those sets. When PLS extracts all mathematically possible components, the resulting linear equation is identical to that produced by MR using all possible independent variables. However, often the crossvalidation criterion limits the number of components to less than the maximum.

Put differently and very roughly, each step (component extraction) of PLS may be described as trying to correlate the dependent variable(s) with **all** the independent variables at once, in contrast with stepwise MR, which at each step identifies the **one** independent variable that best correlates with the dependent variable.

Given this understanding of the PLS algorithm, one can imagine that there are two competing factors that determine whether or not a correlation is observed by PLS:

1) The strength of the "signal", i.e., the proportion of the variance in all of the independent variables together which correlates with variance in the dependent variable(s);
2) The strength of the "noise", i.e., the proportion of the variance in all of the independent variables together which does **not** correlate with variance in the dependent variable(s).

In qualitative terms, PLS will succeed when the overall signal-to-noise ratio is high enough but fail when it is too low.

And in fact this is actually what was observed in the experiments summarized in Tables 3 and 4. As the "noise" resulting from additional random numbers increases, the probability that PLS detects the correlation which was intentionally always present steadily decreased.

Why does the maximum created by these two opposing trends occur within purely random numbers when the numbers of independent variables equals the number of rows? Here we can offer only a very tentative direction for further investigation. It is perhaps significant that this maximum is also the smallest number of variables for which a perfect solution can always be obtained. In classical numerical analysis terms, this situation is equivalent to a system of *n* linear equations in *n* unknowns, which is exactly soluble unless the system is degenerate (*i.e.*, some of the independent variables are exactly linear interdependent, which for random numbers is very unlikely). Until this point, adding more independent variables of random numbers may increase the number of ways in which PLS might find a correlation with the dependent variable(s), just as in stepwise MR each new independent variable of random numbers represents a new opportunity for a correlation with the dependent variable. Beyond this point, the situation for stepwise MR does not change. But with PLS things are more subtle. Each new independent variable does still increase the number of possible unique solutions, very rapidly in fact because when, say, the fifth independent variable is added to a four-row table, there are suddenly five perfect solutions (involving independent variable sets 1,2,3,4; 1,2,3,5; 1,2,4,5; 1.3.4.5; and 2,3,4,5) instead of only one. Perhaps it is this very rapid increase in the possible number of solutions which constitutes the amplified "noise" that so strongly lowers the frequency of chance correlation reported by PLS.

In any case, these findings have very important implications for the practical data analyst, regardless of whether their cause is well understood theoretically. Heretofore the danger of chance correlation with stepwise MR, and the resulting need for "many more rows than independent variables", has inhibited the intellectual development of QSAR. But we have shown above that, with PLS, chance correlation decreases as the number of irrelevant variables increases beyond the number of rows – at least with random numbers, and, from our CoMFA results, probably with correlated variables as well. In short, with PLS one can use as many independent variables as one likes, without increasing the danger of chance correlation. Perhaps CoMFA will be only the first of many new QSAR approaches which takes advantage of these differences in behavior between stepwise MR and PLS.

The lower risk of chance correlation with PLS might seem to be offset by the greater risk of overlooking a "true" correlation, which we found to occur when that correlation involved a sufficiently small fraction of the total variance among the independent variables. As Table 3 shows, if a single column among a hundred irrelevant columns highly correlates with the dependent variable, PLS will probably not report that "fact", whereas stepwise MR certainly will. However, Topliss et al. have already shown that such a "fact" would actually have a high risk of being a chance correlation. Of what practical use could such a dubious "fact" be? Nevertheless, this loss of sensitivity with PLS when "noise" is introduced into a "true" correlation does mean that the analyst should resist the temptation to add independent variables mindlessly. True, the risk of chance correlation does not increase, but the increased risk of overlooking a "real" correlation must also be recognized.

In practice, the tendency for real sets of independent variable data to be intercorrelated reduces the possibility that PLS will overlook a consistent relationship between structure and activity.

Finally, in the specific case of CoMFA, its practitioners should be greatly encouraged by the very low frequency of chance correlation found with molecular fields as the independent variables. In 400 trials with a data set of 21 compounds, Table 4 shows that any r^2_{CV} of 0.25 or greater was found to be significant "at the 95% level", and an r^2_{CV} value of 0.50 or greater is significant "at the 99% level". With as many as 74 compounds, for significance, even at the 99% level, an r^2_{CV} need be little more than positive.

The tendency for PLS to be misled as irrelevant descriptors are added and the signal-to-noise ratio within the independent variables decreases is probably the reason for a CoMFA anomaly that concerns some researchers. CoMFA descriptors are field strength values evaluated on a 3-dimensional rectilinear lattice. It is generally expected that if the lattice spacing is decreased, the r^2_{CV} value should increase. However it more often happens that a rather coarse 2 Angstrom grid spacing produces a somewhat higher r^2_{CV} than does a 1 Angstrom spacing. The current results suggest a reason for this anomaly. Because most regions of 3D space have little relevance to the observed differences in biological activity, a uniform increase in lattice spacing increases the number of noise descriptors much more rapidly than those of signal, and so the r^2_{CV} value is degraded. If this hypothesis is correct, better results should be obtained if lattice spacing is decreased only in those 3D regions which are found relevant to the biological activity [14]. Preliminary studies with this strategy do indeed seem effective in improving r^2_{CV} from CoMFA studies.

In the more general context of computer-aided drug design, these results confirm the growing belief that CoMFA provides one test of a proposed "alignment rule", that is, a procedure for selecting conformations and orientations among a set of biological ligands. If the CoMFA r^2_{CV} value is not positive, the ligands are not all aligned satisfactorily. These results suggest that the converse is also true; if the CoMFA r^2_{CV} value is positive, the alignment is a satisfactory one. However, there is no reason to believe that there is only one alignment rule which will produce a positive r^2_{CV} value from a particular set of ligands, or even that the highest r^2_{CV} value corresponds to the alignments most likely to actually bind to the receptor.

5 Conclusion

Perhaps paradoxically, CoMFA (or any QSAR approach based on the application of PLS with crossvalidation to data tables containing many more columns than rows) is extraordinarily trustworthy. With data sets of a dozen or more compounds, virtually any r^2_{CV} value greater than 0.25 from CoMFA can be accepted as very unlikely to have resulted from chance correlation. However, there is instead some possibility that CoMFA and similar PLS-based variable-intensive approaches can overlook a "true" correlation within a set of data.

6 References

[1] Topliss, J. G. and Costello, R. J., *J. Med. Chem. 15,* 1066–1068 (1971).

[2] Topliss, J. G. and Edwards, R. P., *J. Med. Chem. 22,* 1238–1244 (1979).

[3] Klopman, G. and Kalos, A. N., *J. Comp. Chem. 6,* 492–506 (1985).

[4] Stahle, L. and Wold, S., *Prog. Med. Chem. 25,* 292–337 (1988).

[5] Cramer, R. D. III, Bunce, J. D., Patterson, D. E. and Frank, I., *Quant. Struct.-Act. Relat. 7,* 18–25 (1988).

[6] We thank D. J. Livingston (Smith Kline-Beecham, Welwyn) for calling our attention to this possibility.

[7] Cramer, R. D. III, Patterson, D. E. and Bunce, J. D., *J. Am. Chem. Soc. 110,* 5959–5967 (1988).

[8] Clark, M., Cramer, R. D. III, Jones, D. M., Patterson, D. E. and Simeroth, P. E., *Tetrahedron Comp. Meth. 3,* 47–59 (1990).

[9] Allen, M. S., Tan, Y.-C., Trudell, M. L., Narayanan, K., Schindler, L. R., Martin, M. J., Schultz, C., Hagen, T. J., Koehler, K. F., Codding, P. W., Skolnick, P. and Cook, J. M., *J. Med. Chem. 33,* 2343–2357 (1990).

[10] Mitsutake, K., Iwamura, H., Shimizu, R. and Fujita, T., *J. Agr. Food. Chem. 34,* 725–732 (1986) describes input data; CoMFA study to be published.

[11] The random number generator was the library routine *rand()* supplied by Silicon Graphics.

[12] Wakeling, I. N. and Morris, J. J., submitted to Anal. Chem. We thank these authors for making their manuscript available to us prior to publication.

[13] For an interesting example of this phenomenon within a set of chemical data, see Cramer, R. D. III, *J. Am. Chem. Soc. 102,* 1837 (1980).

[14] Visiting scientist Peter Hecht (Sandoz, Vienna) initially advocated this strategy.

Received on September 20th, 1992; accepted on January 5th, 1993.