QSAR in grossly underdetermined systems: Opportunities and issues

by D. E. Platt
L. Parida
Y. Gao
A. Floratos
I. Rigoutsos

Regression in grossly underdetermined systems has emerged as an important means for understanding molecular activity via comparative molecular field analysis (CoMFA) and other quantitative structure activity relationship (QSAR) studies. But this methodology has applications in much broader areas; for example, near-infrared spectroscopy, mutational enzyme activity studies including protein folding rates to determine which sites are important for determining conformation, and analyses of gene expression data from chip arrays. An error analysis which answers questions concerning the quality of the predictivity, the relative importance of each descriptor, the quality of the estimates of the contribution by each descriptor, and the number of independent components expressed by the associated data is indispensable in understanding whether some particular set of structure variables is important in defining the mechanisms driving the chemical or biological activities. This paper reviews opportunities for QSAR studies. It also considers the analytical aspects of error analysis in least-squares regression, and contrasts principal component regression (PCR) and partial least-squares (PLS) procedures with cross-validation on the issues of error analysis (e.g., the quality of the contribution estimates for each structure descriptor). Further, a methodology for selecting optimal subsets of components in PCR is presented.

1. Introduction

In chemistry, linear regression [1] has provided means for identifying which structural features may be important in determining chemical activity over a group of reactants in a poorly understood interaction. This is achieved by forming a linear relationship between those variables that describe the structural variation within the group of reactants and those that describe the activities of the reactants. The relationship is denoted as a quantitative structure activity relationship (QSAR). QSAR studies have therefore formed a close association with combinatorial chemistry studies in which variations in activities caused

©**Copyright** 2001 by International Business Machines Corporation. Copying in printed form for private use is permitted without payment of royalty provided that (1) each reproduction is done without alteration and (2) the *Journal* reference and IBM copyright notice are included on the first page. The title and abstract, but no other portions, of this paper may be copied or distributed royalty free without further permission by computer-based and other information-service systems. Permission to *republish* any other portion of this paper must be obtained from the Editor.

0018-8646/01/$5.00 © 2001 IBM

by the systematic modifications of structures can yield insight into the reaction activity mechanism.

A particularly interesting example of such a QSAR is comparative molecular field analysis (CoMFA) [2], in which reactants with common structural backbones varying by residue substitutions may be aligned with one another, and physical characteristics such as electrostatic potential and steric energies may be measured for each reactant on a common grid with thousands of points. This is doubly interesting because it explicitly uses descriptors that depend only on the 3D character of a molecule rather than any information on its 1D topology.

CoMFA QSAR regression systems are grossly underdetermined. However, it should be expected that the variations at the various field points should be correlated well enough (because of the relatively small number of residue substitutions as well as the discrete character of residue substitution in combinatoric studies) that a meaningful relationship between those residue substitutions and the activities may be determined. Then those grid points around the residue sites that are important to the determination of the activity will make large contributions to the regression.

One procedure that treats such a grossly underdetermined system is the PLS procedure [3], which was popularized in the CoMFA program of the TRIPOS SYBYL package.[1] The procedure has, since this popularization, emerged as a standard of analysis in numerous research publications. The first applications of CoMFA have also emerged as benchmarks by which other 3D QSARs are measured [2, 4–12].

The application of QSARs to spectral data has a character similar to that of CoMFA. In this case, the spectra are sampled on a one-dimensional lattice of wavelengths. The amplitudes become structure variables for QSAR computations. The application of QSARs to spectroscopy, including near-infrared spectroscopy, shares the overdetermined character of CoMFA studies [13] with a good review by Faber and Kowalski [14].

QSARs have been applied to gene expression analysis in determining levels of gene expression as a function of descriptors of pharmacological substrates or treatment levels [15]. One particular application of regression of particular interest to gene expression studies in general involves the exploration of the relationship between transcriptional and translational control of gene expression [16]. One future application may be predicting survival of alleles, such as cancer survival rates [17], or perhaps of fermentation by yeast alleles [18] as a function of the levels of expression measured by gene array chips.

[1] Available from TRIPOS Associates Inc., 1699 S. Hanley Rd., St. Louis, MO 63144.

Yet another area of increasing interest is that of quantitative structure–property relationships (QSPRs). This is the prediction of physical characteristics such as boiling points, vapor pressure, critical temperature, critical micelle concentrations, and polymer–glass transitions [19]. A possible candidate for application could be the phenomenological exploration of protein folding times. Enzyme mechanism is often elucidated by the examination of mutation activities, just as in combinatoric chemistry studies. Such studies have also been performed on protein folding rates [20]. This form of study is very consistent with standard combinatoric QSAR studies, and represents an opportunity for exploration by QSAR techniques.

But while PLS is a computationally easy method of computing regression coefficients for systems with large numbers of independent variables, the computation of the coefficients is nonlinear in the dependent variable. This implies that error propagation may be performed as a linearized variation, reviewed by Faber and Kowalski [14]. This is frustrating to researchers because the quality of the predictions becomes more difficult to understand. Such an understanding should include the following:

- The range of variation in the regression coefficients that would be consistent with the data.
- The consistency or predictive power of the model with the data as measured by the ability of the structure variables to predict the activities.
- The number of components in the independent variables that actually carry statistical predictive power.
- The contribution of each descriptor to the activity.

One solution to this problem that is also incorporated into SYBYL is cross-validation [21, 22], in which multiple regressions are performed on datasets with various data points excluded from each regression. The variations between regressions for predicted coefficients yield a measure of the range of variation that is consistent with the data, and the quality of the regression is measured by the predicted vs. expected error squared and summed (PRESS) of the activities compared to the variance of the activities. (See Appendix A.) This compares the regression model to a prediction by constant value. PLS does not directly determine regression coefficients. Instead, it determines coefficients for each of a sequence of PLS components. Each of the subsequent PLS components is essentially the correlation between the dependent and independent variables, after the previous component has been projected out of the data. (See Appendix B.) This stops when all of the projections are zero, which happens when, or before, the number of points is reached, or the number of independent variables is reached (whichever is smaller). The number of PLS components that produces the best cross-validation is not usually the one in which

all of the components are projected out. Instead, the best cross-validation occurs at some smaller number of components that represents some measure of the actual number of independent variables contributing to the prediction of the data.

PCR procedures [23, 24] emerge naturally from the quadratic structure of the least-squares problem. The expression for the sum of the squares of the error between predicted and expected values may be expressed as a quadratic form in the regression coefficients. The principal components are orthogonal combinations of the data that diagonalize the coefficient quadratic form. Error propagation in PCR is straightforward and well understood [23, 24]. Further, a least-squares χ^2 statistic to provide a measure of the goodness of fit based on a probability model is commonly used [24–26]. The probability model in minimum χ^2 estimation assumes that the random deviations of the data from the linear regression model are Gaussian.

There are two reasons why PCR has not been adapted for use in CoMFA and other QSAR packages and studies. Most important is that PCR requires the diagonalization of a large matrix. If the number of sampled grid points is in the thousands, the matrix to be diagonalized has a number of components in the millions.

The second reason is a little more subtle. Cross-validation techniques have shown that the best predictivity is not achieved with the inclusion of all of the components, but is usually achieved with some subset. PLS provides only one order of extracted components. However, if there are N independent PCR components, there are 2^N possible subsets that can be constructed from them. This can be prohibitively large to compute. One of the major problems in principal component regression is the selection of an appropriate subset of components. The solution presented by this paper is finding a set of components that extremizes the χ^2 probability, as is typical of some other solutions [23].

There is a difference between cross-validation as a measure of predictivity and χ^2 as a measure of goodness of fit. Goodness of fit measures the consistency of a regression model with data. Cross-validation predictivity measures the extent to which a training set provides information that can predict other unknown data. Practically, this means that goodness of fit includes all points in the set for computing the errors, and cross-validation excludes each point from the prediction computation. In any large dataset, the effect of including or excluding any one point should be diluted. However, the difference can be greatly magnified in any small QSAR set or grossly underdetermined dataset.

Modeling studies, either by linear regression or by more flexible techniques, generally fall into one of two non-exclusive groups: those seeking to understand and measure the contribution of some particular independent or control variables to some putative dependent variable, and those seeking to simply predict the dependent variable as well as possible given some posited control variables, but not necessarily very concerned with the structural consistency of the model itself with the data. Obviously these considerations are not independent or orthogonal, but their interests are sufficiently distinct that the preferred measures of performance tend to be different. Those who are primarily interested in the consistency of a model with the data, or who are interested in determining what the model parameters are, how stable they are, and what their uncertainties are will be more interested in goodness-of-fit considerations. Those who are more interested in prediction will tend to be more interested in predictivity measures and techniques such as bootstrapping, cross-validation, jackknifing, and resampling [27]. This distinction becomes more evident when the researchers interested in determining the parameters are more than willing to confront the complications associated with untangling the impact of correlations between the independent variables on the regression coefficients, whereas those interested in prediction may consider such questions to be unnecessary complications that do not add to understanding.

This paper was motivated by the growing opportunities for practice of the application of linear regression, particularly in grossly overdetermined systems, to biological problems, which were reviewed earlier in this section. However, the focus is on those problems associated with regression that are peculiar to the grossly overdetermined systems that are emerging in those biological applications, with an emphasis on the measurement of regression coefficients and goodness of fit.

2. Regression model

The regression model is expressed simply as

$$y_i = \sum_j x_{ij} a_j + e_i , \tag{1}$$

where y_i is the ith measurement out of N of the dependent variable (activity for the ith molecule in the dataset), x_{ij} is the ith measurement of the jth out of D independent variables (jth QSAR structure descriptor for the ith molecule in the dataset), a_j is the jth regression coefficient, and e_i is the error in the prediction of the ith dependent variable by the regression. Many regression studies use Greek letters to represent estimated regression parameters, but this usage is not universal [24, 26]. The expected values of e_i are described as

$$E(e_i) = 0, \tag{2}$$

$$E(e_i e_j) = \Delta y_i^2 \delta_{ij} . \tag{3}$$

If each of the errors e_i is Gaussian distributed, the statistic

$$\mathfrak{S}^2 = \sum_i \frac{e_i^2}{\Delta y_i^2} = \sum_i \frac{1}{\Delta y_i^2}\left(y_i - \sum_j x_{ij} a_j\right)^2 \tag{4}$$

is χ^2 distributed with N degrees of freedom [23–26]. This minimization of $\mathfrak{S}^2$ is the foundation of both PLS and PCR. It may be expressed in terms of matrices as

$$\mathfrak{S}^2 = (Xa - y)^T C(Xa - y), \tag{5}$$

where C is the diagonal matrix with elements $(C)_{ij} = \delta_{ij}/\Delta y_i^2$. This may be expressed alternatively as

$$\mathfrak{S}^2 = (a - a_0)^T X^T CX(a - a_0) + y^T Cy - a_0^T X^T CXa_0, \tag{6}$$

where

$$a_0 = \lim_{\epsilon \to 0}(X^T CX + \epsilon I)^{-1} X^T Cy. \tag{7}$$

The limit $\lim_{\epsilon\to 0}(X^T CX + \epsilon I)^{-1} X^T C^{1/2}$ is called a "generalized inverse" of $C^{1/2}X$. The limit $\lim_{\epsilon\to 0}(X^T CX + \epsilon I)^{-1}$ is undefined unless the matrix first operates on another matrix or vector which has no projections along eigenvectors of $X^T CX$ that correspond to eigenvalues equal to zero. However, if u is an eigenvector of $X^T CX$ with a zero eigenvalue $(Xu)^T C(Xu)$, it follows that any projection Xu of X along u will be zero, since $X^T CXu = 0$ implies that $u^T X^T CXu = (Xu)^T C(Xu) = 0$. Since C is diagonal, this implies that each $(Xu)_i = 0$. Further, this implies that $u^T a_0 = 0$. Note that this solution is not unique. Any $a_0' = a_0 + \delta a$, where $X^T CX\delta a = 0$, produces an equivalent $\mathfrak{S}^2$.

Since $\mathfrak{S}^2$ is a χ^2 statistic, it follows that a_0 and $X^T CX$ are essential statistics. Any changes in a may be accounted for by the contribution to the error $\mathfrak{S}^2$ by the coefficients

$$\mathfrak{S}^2_{\text{coef}} = (a - a_0)^T X^T CX(a - a_0), \tag{8}$$

with the remainder accounted for by the residual

$$\mathfrak{S}^2_{\text{res}} = y^T Cy - a_0^T X^T CXa_0, \tag{9}$$

so that

$$\mathfrak{S}^2 = \mathfrak{S}^2_{\text{coef}} + \mathfrak{S}^2_{\text{res}}. \tag{10}$$

The total number of degrees of freedom in $\mathfrak{S}^2$ is N. The number of degrees of freedom in $\mathfrak{S}^2_{\text{coef}}$ is equal to the number D_0 of eigenvectors with corresponding nonzero eigenvalues of $X^T CX$. This leaves $N - D_0$ degrees of freedom for $\mathfrak{S}^2_{\text{res}}$. This partition is very reminiscent of Bayesian treatments of linear regression [28], but the presentation here follows "frequentist" notions of sampling theory.

The expectation value of a_0 is

$$E(a_0) = E\left\{\lim_{\epsilon\to 0}(X^T CX + \epsilon I)^{-1} X^T Cy\right\} = a, \tag{11}$$

where a is the fixed parameter characterizing sample space, to within the previously described ambiguity. Both PLS and PCR solve the same minimization of $\mathfrak{S}^2$ if all components are employed, and so both obtain the same value of a_0. The covariance is predicted by

$$\text{cov}(a_0, a_0) = E[(a - a_0)(a - a_0)^T] = \lim_{\epsilon\to 0}(X^T CX + \epsilon I)^{-1}, \tag{12}$$

determined by the inverse of the quadratic coefficients in $\mathfrak{S}^2_{\text{coef}}$. As pointed out before, this limit does not exist if there are eigenvalues of $X^T CX$ equal to zero. This means that any contribution to a of any magnitude in a direction corresponding to a null eigenvector of $X^T CX$ does not contribute anything to $\mathfrak{S}^2$, and implies that the coefficients essentially have an infinite uncertainty and are completely undetermined in any underdetermined system. This is just another reflection of the ambiguity in underdetermined systems.

A meaningful alternative measure of covariance is the amount by which the estimate of a will vary given the variations in y. This is essentially equivalent to the effect of allowing y to vary according to the variation in e, implying that

$$\begin{aligned}
\text{cov}_{\text{subspace}}(a_0, a_0) &= E\{a_0[e]a_0[e]^T\} \\
&= \lim_{\epsilon\to 0} E\{(X^T CX + \epsilon I)^{-1} X^T Cee^T CX(X^T CX \\
&\quad + \epsilon I)^{-1}\} \\
&= \lim_{\epsilon\to 0}(X^T CX + \epsilon I)^{-1} X^T CE(ee^T)CX(X^T CX \\
&\quad + \epsilon I)^{-1} \\
&= \lim_{\epsilon\to 0}(X^T CX + \epsilon I)^{-1} X^T CC^{-1} CX(X^T CX \\
&\quad + \epsilon I)^{-1},
\end{aligned}$$

or

$$\text{cov}_{\text{subspace}}(a, a) = \lim_{\epsilon\to 0}(X^T CX + \epsilon I)^{-1} X^T CX(X^T CX + \epsilon I)^{-1}. \tag{13}$$

The limit does exist in this case because $(X^T CX + \epsilon I)^{-1} X^T CX$ acts like a projection operator that picks out only those eigenvectors with nonzero eigenvalues. This expression compares favorably with the variation in the coefficients observed between the various regressions produced by cross-validation. Such a result constitutes

an explicit measure of the stability of the coefficients to variations in the dependent variables.

It is important to realize that while some consistency may be expected within a dataset, and it is possible to ask whether a model is consistent with a dataset in a statistical sense, underdetermined systems do not yield definitive measurements of all of the coefficients. Comparison with other datasets that could ultimately produce a complete model if the data were combined would not produce coefficients consistent with one another.

Not only is it possible for a regression to be underdetermined in the sense of having eigenvalues equal to zero, but some of the eigenvalues of X^TCX may be very small. Such a system is called "poorly conditioned." This corresponds to some $\mathrm{var}(a_i)$ being very large. Such terms can add spurious and large contributions to a_0 without significantly affecting $\mathfrak{E}^2$, suggesting that it is desirable to exclude contributions from various subsets of components that may not correspond to zero-valued eigenvalues. The systematic consideration of the character of individual principal components in the analysis of the $\mathfrak{E}^2$ quadratic form is perhaps the best definition for principal component regression. This includes but is not limited to issues of conditioning of the regression equations. While PCR provides a direct way to examine questions such as the conditioning of a regression, PLS provides no direct way to consider the issue of whether a regression is poorly conditioned or not.

Consider a projection operator P that is a projection onto a subset K of eigenvectors u_k of X^TCX. As such, P satisfies

$$P = \sum_{k \in K} u_k u_k^T, \tag{14}$$

$$P^2 = P, \tag{15}$$

$$[P, X^TCX] = 0. \tag{16}$$

Now, it is desirable to partition $\mathfrak{E}^2$ in a different way along projections of Pa:

$$\mathfrak{E}^2[P] = (Pa - a_0)^T X^T CX(Pa - a_0) + y^T Cy - a_0^T X^T CXa_0 .$$

Identifying the projection operator $Q = I - P$, and noting that

$$(Pa - a_0)^T X^T CX(Pa - a_0) = (Pa - Pa_0)^T X^T CX(Pa - Pa_0) + (Qa_0)^T X^T CX(Qa_0),$$

and that

$$(a_0)^T X^T CX(a_0) = (Pa_0)^T X^T CX(Pa_0) + (Qa_0)^T X^T CX(Qa_0),$$

it follows that

$$\mathfrak{E}^2[P] = \mathfrak{E}^2_{\mathrm{coef}}[P] + \mathfrak{E}^2_{\mathrm{res}}[P], \tag{17}$$

where

$$\mathfrak{E}^2_{\mathrm{coef}}[P] = (Pa - Pa_0)^T PX^T CXP(Pa - Pa_0) \tag{18}$$

and

$$\mathfrak{E}^2_{\mathrm{res}}[P] = y^T Cy - a_0^T PX^T CXPa_0 . \tag{19}$$

This is a very interesting partition of the degrees of freedom. The operator P removes degrees of freedom from $\mathfrak{E}^2_{\mathrm{coef}}[P]$ and essentially transfers them to $\mathfrak{E}^2_{\mathrm{res}}[P]$. Since the total number of degrees of freedom in $\mathfrak{E}^2[P]$ remains N, and the number of degrees of freedom in $\mathfrak{E}^2_{\mathrm{coef}}[P]$ is now D_P, the total number of degrees of freedom in $\mathfrak{E}^2_{\mathrm{res}}[P]$ is now $N - D_P$. Further, while the number of degrees of freedom in $\mathfrak{E}^2_{\mathrm{res}}[P]$ increases when poorly conditioned eigenvectors are excluded, so does the total value of $\mathfrak{E}^2_{\mathrm{res}}[P]$. The relationship between the goodness-of-fit error and the exclusion of particular components is well understood [23] and has been considered in comparisons between PLS and PCR [13, 14].

Partitioning the error according to contributions by projections of components has an immediate application, allowing comparison of the goodness of fit for different subsets of components. In particular, for a subset of components projected by P, the probability that a χ^2 larger than this might be observed is $P(\chi^2_{N-D_P} > \mathfrak{E}^2_{\mathrm{res}}[P])$. Those with larger probabilities better represent the fit. Note that if $N = D_P$, which happens when all of the non-null components are used in an underdetermined system, $P(\chi^2_{N-D_P} > \mathfrak{E}^2_{\mathrm{res}}[P])$ is undefined. There is essentially no statistical information about the quality of the fit if all of the principal components are included.

Further, the contributions of each individual component may also be determined. The contribution to $\mathfrak{E}^2_{\mathrm{res}}[P]$ may be determined for any particular component k. For any component k with eigenvector u_k, the projection operator is $P_k = u_k u_k^T$. This implies that the effect of any particular eigenvector is to subtract a variation

$$\mathfrak{E}^2_k = a_0^T u_k u_k^T X^T CXu_k u_k^T a_0$$
$$= y^T CX(X^T CX + \epsilon I)^{-1} u_k u_k^T X^T CXu_k u_k^T (X^T CX + \epsilon I)^{-1} X^T Cy$$

or

$$\mathfrak{E}^2_k = \frac{y^T C(Xu_k)(Xu_k)^T Cy}{(Xu_k)^T C(Xu_k)}, \tag{20}$$

where $A_k = (Xu_k)^T C(Xu_k)$ is the eigenvalue of X^TCX corresponding to eigenvector u_k. The contribution of the kth component to $\mathrm{cov}(a, a)$ varies as $1/A_k$. This is a reflection of how well conditioned the contribution from this component is. Small A_k components contain little discriminating information compared to the uncertainty they contribute to the regression coefficients. Exclusion of the smallest A_k contributions therefore improves the

stability of the coefficients and reduces the size of the uncertainty in those parameters.

However, the largest $\mathfrak{S}_k^2$ contribute the most toward improving the goodness of fit, since they reduce $\mathfrak{S}_{\text{res}}^2[P]$ the most. Therefore, the value of $\mathfrak{S}_k^2$ represents the predictive power of the kth component. It is possible therefore to rank the components by predictive power. Consequently, it is possible to construct a list of subsets with the largest predictive power, then the next list containing the largest together with the second largest, and then the third list containing the top three predictive components, etc. This reduces the computation from all 2^N possible subsets of components to a simple list of subsets N long. Once this is done, it is possible to compute $P(\chi^2_{N-D_P} > \mathfrak{S}^2_{\text{res}}[P])$ for each of the subsets. This probability generally passes through some extremum, which represents the optimal subset of components. Since the questions of the information in a component as measured by A_k and the contribution the component makes to the goodness of fit are distinct, exclusion of low-information components may be achieved by applying a cutoff to A_k. A selection of the most important contributors to the goodness of fit may then be applied.

This approach may be applied to the situation in which the size of the uncertainty is unknown and it is desired to estimate some best uncertainty from the regression of the data. This may be achieved by choosing $C = I/\Delta Y^2$, to yield

$$E(\mathfrak{S}^2_{\text{res}}[P]) = N - D_P = \frac{1}{\Delta Y^2}\,(y^T y - a_0^T P X^T X P a_0), \tag{21}$$

and solving for ΔY^2. The best subset is the one that produces the smallest ΔY^2. This component-selection criterion is essentially identical to one proposed by Lott [29], who also recognized the possibility of reducing the optimal space of subsets by ranking the components. However, the connection between the selection of an optimal subset and a minimum ΔY^2 was not established, and connection with χ^2 was not explored. Generally, for overdetermined systems, ΔY goes through a minimum as the number of components is decreased. The smallest set is the best. However, in underdetermined systems there tends to be no minimum in ΔY. For a fixed ΔY, there is usually some particular subset of components where $P(\chi^2_{N-D_P} > \mathfrak{S}^2_{\text{res}}[P])$ is minimized. Once some ΔY is selected and the component subset is extracted, the values of a_0 and cov(a, a) that are consistent with the quality of the regression and the variation in the data may be computed.

Principal component regression follows the plan of principal component analysis (PCA): It is assumed that the solution space is best represented by some subset of components. But the problem of PCA is to find a subset of components that represents the data to within some limit of accuracy. In the case of PCA, the most important components are those with the largest variance. (See Appendix C.) The problem with selecting a subset that spans the space of variation in PCR is that the dependent variable may depend on some of the components with smaller variation: If the short axis is discarded, there is no dimension to describe the layering of its structure. This problem is well known, and there have been a number of solutions posed for selecting some optimal subset of components [23]. However, many commercial packages still rank components according to their variation $A_k = (Xu_k)^T C (Xu_k)$, which measures only the information in the component, and *not* the contribution of the component $\mathfrak{S}_k$ to the goodness of fit. This has led to some unfounded criticisms of PCR in various comparisons with PLS. The use of $P(\chi^2_{N-D_P} > \mathfrak{S}^2_{\text{res}}[P])$ as the measure of the quality of fit for component subsets presented here appears to be a new contribution.

One of the prohibitive costs of PCR computation is the diagonalization of the large matrix. However, since small A_ks yield poorly conditioned regressions, and zero-valued A_ks correspond to undetermined contributions not spanned by the dataset, it is appropriate to exclude them, as in PCA. This implies that a computational algorithm such as NIPALS [30] may be applied, which is much more efficient if only the first few components are desired.

3. Conclusions

Partial least-squares analysis and cross-validation have emerged as standards in analyzing 3D QSARs because of their simplicity of computation. Together, they provide methods for assessing

- The range of variation in the regression coefficients that would be consistent with the data.
- The consistency or predictive power of the model with the data as measured by the ability of the structure variables to predict the activities.
- The number of components in the independent variables that actually carry statistical predictive power.
- The contribution of each descriptor to the activity.

Since PLS and PCR address these issues differently, it is instructive to compare them. The most significant difference is that the contributions by the individual PCR components are immediately available, and the implications of each one with respect to each of the items in the above list are computable. PLS components are much more intermixed; none of the components have meaning outside of the context of the rest.

First, PLS allows for the examination of the variation between descriptor coefficients by comparing the cross-validation regressions. Usually variances and covariances

in the coefficients are not provided, though estimates could be computed from the variation in cross-validation regressions. In PCR, the contribution of each component to the variances and covariances of the coefficients is immediately available. In either case, the availability of this information permits the assignment of error bars to estimates of activity.

Second, there is a difference between goodness-of-fit measurements as defined by $P(\chi^2_{N-Dp} > \mathfrak{E}^2_{\text{res}}[P])$ and the cross-validation predictivity. On one hand, the predictivity measures how well the set can predict data that are not included in the set. This implies that it measures extrapolative power, while goodness of fit measures the quality of consistency of the data, or the interpolative power. In this sense, at least in nondilute or data sparse sets, predictivity may appear to be more demanding than goodness of fit. However, the reference variation that is used in the predictivity measurement is the variance in the independent variables (activities), while the reference variation in the goodness of fit is the distribution of deviations of the model from the actual data. The errors in the dependent variable should be much more restrictive than the variation in the independent variable. If the error bars are larger than the total range of measured variation, the signal is indistinguishable from noise. Predictivity requires only that the regression perform better than the total variation in the dependent variable. Goodness-of-fit consistency requires that the regression perform better than the uncertainty in the dependent variable. Usually a regression that passes a predictivity test does not pass a consistency test to within experimental error. In most cases, it is better to estimate the error as an unknown, attributing some of its magnitude to variables that have not been accounted for in the regression model. In that case, the estimated variance is usually much smaller than the variance of the independent variables.

Third, the use of cross-validation implies that there is some variation between the regressions. It can be difficult to identify corresponding PCR components between separate cross-validation regressions, especially in grossly underdetermined systems. However, it is generally possible to identify which of the components of leading importance maintained their importance from regression to regression, as well as how much those components vary among regressions. In PLS, it is very difficult to understand the relationships among the components, much less how the decomposition varies among regressions and what that variation in the decomposition might mean. This exposes at least one problem with cross-validation; the regression to some extent compares unrelated components in choosing which models with which components most closely represent the data. The reason is that the least-squares solutions obtained by minimizing $\mathfrak{E}^2$ are not optimizations of predictivity measured by PRESS, since predictivity is based on points not included in the regression and thus not included in $\mathfrak{E}^2$. At least one implication of this fact is that it is difficult to assess the impact of excluded data on the regression coefficients analytically. The decomposition of PLS selects those parts of the independent variation that correlate most strongly with the dependent variations at each step. Usually the number of PLS components appears to be less than the number of PCR components. However, those components are distinct in their character and should not be compared directly with one another. Usually regressions requiring more PLS components than other regressions also require more PCR components than the others.

Fourth, the relative contribution of each of the descriptors to the variance of the dependent variable is presented in SYBYL as $|a_i|\sqrt{\text{var}(x_i)/\text{var}(y)}$. However, this does not take into account the effects of correlation among the descriptors within the sampled set. Further, this computation is also available within the context of PCR, and PCR allows for the identification of the importance of individual components to each regression by examining the projections Xu_k for each principal component k, taken together with the predictive power of the component, as well as the relative contributions (amplitudes) of the descriptors to the components in u_k.

Perhaps most important, it is possible to recognize when extrapolation data points contain information outside the subspace spanned by the data. For example, if u_k corresponds to a zero eigenvalue, $x^T u_k$ is a measure of the projection of that extrapolation candidate into uncharacterized space, where the contribution to the activity will be undefined. This information is not as directly available in PLS, even though it is possible to examine the residual in the independent variables after decomposition by the training set has been performed.

More, if u_k corresponds to a poorly conditioned component, $x^T u_k$ measures the projection into badly characterized space, whose contribution to the activity is poorly defined. This information is not available in PLS, and care must be exercised to recognize when spurious results follow from data that are poorly conditioned. A good example of this would be the inclusion of several descriptors which are constrained to sum to zero. PLS may recognize the round-off error as a part of the regression, yielding large coefficients for those descriptors.

One argument against PCR is that the components are an artifact of the space that happens to be sampled by the dataset. For example, the components themselves depend on how the independent variables were scaled, and the quality and predictive power of the components can be changed merely by changing the units of measurement. This shows up particularly in that measurement unit changes can change the activity predictions in using some number K of components. It is particularly true in the case

of underdetermined systems, in which strong correlations between the descriptors may be induced simply by the small sample size compared to the number of descriptors. Such pathology happens to be true of PLS as well, but PLS hides its appearance in its decompositions more effectively. The technique most often adapted by both PCR and PLS is to rescale the descriptors by the root of the variance of that descriptor within the dataset; this tends to give a more balanced variation between descriptors.

Ultimately PCR and PLS are founded on the minimization of a sum-of-squares error expression. This paper has explored some of the analytic consequences of least-squares solutions as they apply to PLS and PCR, and has considered how they address those issues of error analysis that have emerged as being very important to QSAR studies.

Appendix A: Cross-validation

Since the number of degrees of freedom associated with individual PLS components is not known, the construction of a χ^2 error associated with a truncated set of components is not meaningful. One way around that is to ask how well the individual, and it is assumed independent, points are predicted by a fit to all of the other points. This leads to the idea of cross-checking the fits of all of the data points via a method called "cross-validation." Further, it is desirable to try to determine which components, or how many, are necessary to predict the data. By applying cross-validation to predictions using varying numbers of components, it is possible to determine the number of components that produces the best predictive capability. This approach partly motivated the development, presented in this paper, of a PCR component subset selection based on χ^2 contributions that measured the probability that a linear model could describe a set of data consistent with errors of measurement.

Cross-validation is a technique that may be applied either to PLS or to PCR. The method cycles through the set of points, excluding one point from the computation of the fit parameters, and then constructing an error for the fit vs. actual excluded point. A sum-of-squares error is constructed for this and is used in an F-test to compare against the variance about the mean. If this were applied to PCR, it is not clear how many degrees of freedom should be assigned to the number of independent components that were kept for the fit, since each component varies from point to point for each of the predicted points that was excluded from the computation for that fit. For PLS, where each retained component does not contribute a χ^2 degree of freedom in the first place, it is even more unclear how many degrees of freedom to assign to each retained component. However, common practice seems to be to assign one degree of freedom per component.

The statistics that apply follow. The predicted sum of squares PRESS is defined as

$$\text{PRESS} = \sum_i [y_i - {}_n\hat{y}_i]^2,$$

where ${}_n\hat{y}_i$ is the predicted value of y for the ith point using n PLS components. A q^2 that is much like a correlation coefficient compares the ratio of PRESS to the variance

$$q^2 = 1 - \frac{\text{PRESS}/N}{\sigma_y^2},$$

where N is the number of points. The F statistic is constructed from

$$F_{N,N} = 1 - q^2 = \frac{\text{PRESS}/N}{\sigma_y^2}.$$

The numerator is essentially the variance from prediction; the denominator is the no-prediction (all values predicted by the mean of the independent variables) variance. If the denominator is a variance estimated about a sample mean, the numerator has a number of degrees of freedom associated with the number of data points N, but the denominator has the degrees of freedom associated with the same number of data points, but estimated about the mean $N - 1$. The actual F statistic should then be

$$F_{N,N-1} = \frac{\text{PRESS}/N}{s^2},$$

where the s^2 has an $N - 1$ in it.

The literature frequently cites q^2 without interpretation in terms of F scores or probability estimates, thus avoiding the question of whether the number of degrees of freedom are defined. Assuming they are defined, the critical value for q^2 given the degrees of freedom has the form

$$q^2 = 1 - F.$$

At the 95% level, if $N = 2$, the value of F is $F = 0.054017$ for a $q^2 = 0.9459$. For $N = 6$, $F = 0.227927$ for a $q^2 = 0.7731$. For $N = 10$, $F = 0.331084$ for a $q^2 = 0.6689$. For $N = 100$, $F = 0.718046$ for a $q^2 = 0.2819$. Depending on the number of data points, the range over which a critical q^2 can vary is significant. Frequently a q^2 is cited in the literature with no sense of whether it reflects a statistically significant level.

Appendix B: Partial least squares

The approach taken here is to iteratively extract components each one of which satisfies the form of the least-squares equation. Each of the components is linearly independent of the previous ones, until no more components can be constructed [3].

In this case, just as in the PCR case, the coefficients a in a vector of length D in an expression of the form

$$x^T a = y_p,$$

for a particular vector x composed of D descriptors and a scalar y_p, are determined by minimizing a χ^2 error,

$$\mathfrak{S}^2 = \sum_{i=1}^{N} w_i (x_i^T a - y_i)^2 = (Xa - y)^T C (Xa - y),$$

where $w_i = 1/\Delta y_i^2$, C is diagonal with $(C)_{ii} = w_i$, y is now a vector N long, and X is an $N \times D$ matrix. The minimization of $\mathfrak{S}^2$ with respect to a yields

$$c \equiv X^T C X a = X^T C y$$

for a vector c of D components. This equation is to be solved for a. One problem is that the matrix multiplying a may be singular. If the goal is to extract the most important factor leading to y, something related to the correlation of the y with X should be examined. This might most simply be done by taking the projection along c, which is just that correlation. Then

$$\|c\| = \hat{c}^T c = \hat{c}^T X^T C y = \hat{c}^T X^T C X a.$$

Defining $u = X\hat{c}$, $u_i = x_i^T \hat{c}$, and identifying

$$b \equiv \frac{u^T C X}{u^T C u},$$

it follows that

$$b\hat{c} = 1$$

and

$$ba = \frac{u^T C y}{u^T C u}.$$

If the preceding describes how the structure components line up with the response or activity data, that information may be used to extract the projection of this dominant leading component. A transformation is sought that subtracts displacements D from X such that $X' = X - D$ are perpendicular to $\hat{c}$. At the same time, there should be a δy which, when subtracted from y, leads to the corresponding $y' = y - \delta y$. The connection between X and y is through a in $y = Xa$, which is to be preserved in the projected components $y' = X'a$ with the same a. This implies that the transformation will involve a $\delta y = Da$. The only vector we have constructed that has a well-defined scalar product relationship with a is b. Therefore, we want to construct D such that $D\hat{c}$ acts like a projection with $b\hat{c}$. This may be done by seeking f such that $D = fb$ and

$$X' = X - fb,$$

where

$$X'\hat{c} = 0.$$

This is satisfied when

$$X'\hat{c} = X\hat{c} - fb\hat{c} = u - f = 0,$$

so

$$f = u$$

and

$$X' = X - ub.$$

Forming the inner product of this with a yields

$$X'a = Xa - uba.$$

Recognizing that $y' = X'a$, $y = Xa$, and $ba = u^T C y / u^T C u$, this simplifies to

$$y' = y - u \frac{u^T C y}{u^T C u}.$$

Once the projected X and y have been obtained, these may be used to compute new c and b, and new projections obtained until $c = 0$. Since each new c is constructed from linear combinations perpendicular to all previous c, each c is perpendicular to all previous c. If fewer data points than components are present, the termination condition is met in a number of iterations smaller than the number of components. The order of subtraction proceeds from largest to smallest correlation with the y.

There is no explicit computation of a as so far defined. Instead, predictions are obtained by decomposing an unknown x by the set of c and b that were determined by the iterative application of the above projections on the training set. Each iteration folds the data back on itself nonlinearly. In order to predict a particular y_p for some x, a reverse application of the above projections must be applied. For each iteration, x must be decomposed using

$$x' = x - vb^T,$$

where $v = x^T \hat{c}$. The remaining x, after the complete decomposition, is a measure of the information in x that was not accounted for by the original dataset that was used to construct the least-squares decomposition. Starting with $y_p = 0$, the new y must be composed,

$$y_p = y_p' + v \frac{u^T C y}{u^T C u},$$

for the b and c corresponding to the step of decomposition that was determined from the previous decomposition of the data.

Now, if

$$y = m_1 x_1 + m_2 x_2 + m_3 x_3 + \cdots,$$

it follows that $x = (1, 0, 0, \cdots)$ will yield $y = m_1$, $x = (0, 1, 0, \cdots)$ will yield $y = m_2, \cdots$, etc.

The relationship between the error $\mathfrak{E}^2$ between steps extracting components may be extracted as follows. First,

$$\begin{aligned}\mathfrak{E}^2 &= (Xa - y)^T C(Xa - y)\\ &= \left(X'a - uba - y' + u\frac{u^TCy}{u^TCu}\right)^T\\ &\quad \cdot C\left(X'a - uba - y' + u\frac{u^TCy}{u^TCu}\right)\\ &= \left(X'a - u\frac{u^TCy}{u^TCu} - y' + u\frac{u^TCy}{u^TCu}\right)^T\\ &\quad \cdot C\left(X'a - u\frac{u^TCy}{u^TCu} - y' + u\frac{u^TCy}{u^TCu}\right)\\ &= (X'a - y')^T C(X'a - y').\end{aligned}$$

Therefore, the extraction of each component leaves the error unchanged. The error for a PLS regression is determined entirely by the residual for the last extracted component in the subset of components. This is entirely different in character from the notion in PCR that each component contributes a predictive capability that can be measured by its "predictive power" (see Section 2).

This method has some advantages. First, it produces leading contributing structure components that predict the strongest correlations with the activities in order. There is no need to compute any more than is desired.

However, there are a number of disadvantages. First, there is no really overt computation for a. Instead, PLS provides a decomposition technique. The dependence of the error $\mathfrak{E}^2$ on a is never explicitly determined. Further, even the contributions of each individual component of $\mathfrak{E}^2$ depend on the entire subset that was applied. The method is nonlinear in the activities, which makes a propagation of errors very difficult. Since the components are not orthogonal in quadratic contributions to the χ^2 statistic $\mathfrak{E}^2$, it is impossible to determine the number of degrees of freedom that have been involved if some components are dropped, and therefore there is no internal goodness-of-fit statistic. Instead, a "cross-validation" technique [21, 22] is usually employed to try to extract this information. Further, the components must be picked off in order. It is impossible to measure the individual contributions of each component outside the decomposition sequence, so an independent articulation of the contribution of each component, as with eigenvectors, is impossible.

In conclusion, there is no direct estimation of a. The nonlinearity in y makes a propagation of errors difficult, and a direct way to compute $\mathrm{cov}(a_i, a_j)$ is unavailable. Since the components are not orthogonal in quadratic contributions to the χ^2 statistic $\mathfrak{E}^2$, it is impossible to determine the number of degrees of freedom contributed by the components; therefore, there is no internal estimate of goodness-of-fit statistic. To paraphrase, while one can determine the simplest elements of the fit, it is not readily clear how much is known, how well it is known, or even whether something new is not known from previous experience.

Appendix C: Principal component analysis

The notion behind principal component analysis [23] is that data may be approximated most effectively by the leading principal components of the covariance matrix. This may be understood as follows. Consider a set of N vectors $\vec{x}_i$, which is represented in terms of variations about the mean

$$\vec{x}_i = \vec{\xi}_i + \vec{X},$$

so that $(1/N)\ \Sigma_i\ \vec{x}_i = \vec{X}$; and whose variation about said mean is expressed in terms of the eigenvectors of the matrix $\mathbf{C} = (1/N)\ \Sigma_i\ \vec{\xi}_i\vec{\xi}_i^T$. These may be written

$$\mathbf{C}\hat{v}_k = \sigma_k^2\hat{v}_k\,.$$

Then, since the $\hat{v}_k$ span the space, the $\vec{\xi}_i$ may be expressed in terms of the $\hat{v}_k$ as

$$\vec{\xi}_i = \sum_k c_{ik}\hat{v}_k\,.$$

Since $\mathbf{C}$ is symmetric with real coefficients, the $\hat{v}_k$ are orthogonal. It follows that

$$c_{ik} = \hat{v}_k^T\vec{\xi}_i$$

and that $(1/N)\ \Sigma_i\ c_{ik}^2 = (1/N)\ \Sigma_i\ \hat{v}_k^T\vec{\xi}_i\vec{\xi}_i^T\hat{v}_k = \hat{v}_k^T\mathbf{C}\hat{v}_k = \sigma_k^2$. If $\sigma_k = 0$, it follows that $c_{ik} = 0$. Thus, the c_{ik} may be expressed as proportional to σ_k, or $c_{ik} = u_{ik}\sigma_k$, so that

$$u_{ik} = \frac{1}{\sigma_k}\hat{v}_k^T\vec{\xi}_i \text{ (where } \sigma_k \neq 0)$$

and

$$\vec{\xi}_i = \sum_{k,\sigma_k\neq 0} \sigma_k u_{ik}\hat{v}_k\,.$$

Then $(1/N)\ \Sigma_i\ u_{ik}^2 = 1$. Next, the question is how much each of the components contributes to the representation of the data. In particular, if some subset S of the components is used, and its complement S' is excluded, how good is the approximation? We define an error

$$\mathcal{E}^2(S) = \frac{1}{N}\sum_i \left|\sum_{k\in S} \sigma_k u_{ik}\hat{v}_k - \vec{\xi}_i\right|^2$$

$$= \frac{1}{N}\sum_i \left|\sum_{k\in S'} \sigma_k u_{ik}\hat{v}_k\right|^2$$

$$= \frac{1}{N}\sum_i \sum_{k,k'\in S'} \sigma_{k'}\sigma_k u_{ik'} u_{ik}\hat{v}_{k'}^T\hat{v}_k$$

$$= \frac{1}{N}\sum_i \sum_{k\in S'} \sigma_k^2 u_{ik}^2$$

$$= \sum_{k\in S'} \sigma_k^2 .$$

The error is then equal to the sum of the eigenvalues of the covariance matrix that were excluded from the approximation. If these are the smallest eigenvalues, the approximation shows a minimum error.

References

1. K. Pearson, "On Lines and Planes of Closest Fit to Systems of Points in Space," *Philosoph. Mag.* **6,** No. 2, 559–572 (1901).
2. R. D. Cramer III, D. E. Patterson, and J. D. Bunce, "Comparative Molecular Field Analysis (CoMFA). Effect of Shape on Binding of Steroids to Carrier Proteins," *J. Amer. Chem. Soc.* **110,** 5959–5967 (1988).
3. H. Wold, "Nonlinear Estimation by Iterative Least Squares Procedures," *Festschrift for J. Neyman*, F. N. David, Ed., John Wiley, New York, 1966; H. Wold and E. Lyttkens, "Nonlinear Iterative Partial Least Squares (NIPALS) Estimation Procedures," *Bull. Intern. Statist. Inst: Proc. 37th Session*, London, 1969, pp. 1–15.
4. K. H. Kim and Y. Martin, "Direct Prediction of Dissociation Constants (pKa's) of Clonidine-Like Imidazoles, 2-Substituted Imidazoles, and 1-Methyl-2-Substituted-Imidazoles from 3D Structures Using a Comparative Molecular Field Analysis (CoMFA) Approach," *J. Med. Chem.* **34,** 2056–2060 (1991).
5. K. H. Kim and C. M. Martin, "Direct Prediction of Linear Free Energy Substituent Effects from 3D Structures Using Comparative Molecular Field Analysis. 1. Electronic Effects of Substituted Benzoic Acids," *J. Org. Chem.* **56,** 2723–2729 (1991).
6. E. A. Coats, "The CoMFA Steroids as a Benchmark Dataset for Development of 3D QSAR Methods," *3D QSAR in Drug Design*, Vol. III, H. Kubinyi, G. Folkers, and Y. C. Martin, Eds., Kluwer Academic Publishers, Dordrecht, The Netherlands, in press, p. 199.
7. R. D. Cramer III, "BC(DEF) Parameters. 1. The Intrinsic Dimensionality of Intermolecular Interactions in the Liquid," *J. Amer. Chem. Soc.* **102,** 1837–1849 (1980).
8. A. C. Good, S. Sung-Sau, and W. G. Richards, "Structure-Activity Relationships from Molecular Similarity Matrices," *J. Med. Chem.* **36,** 433–438 (1993).
9. A. C. Good, S. J. Peterson, and W. G. Richards, "QSAR's from Similarity Matrices. Technique Validation and Application in the Comparison of Different Similarity Evaluation Methods," *J. Med. Chem.* **36,** 2929–2937 (1993).
10. A. N. Jain, K. Koile, and D. Chapman, "Compass: Predicting Biological Activities from Molecular Surface Properties. Performance Comparisons on a Steroid Benchmark," *J. Med. Chem.* **37,** 2315–2327 (1994).
11. M. S. Allen, Y. Tan, M. Ml. Trudell, K. Narayanan, L. R. Schindler, M. J. Martin, C. Schultz, T. J. Hagen, K. F. Koehler, P. W. Codding, P. Skolnick, and J. M. Cook, "Synthetic and Computer-Assisted Analyses of the Pharmacophore for the Benzodiazepine Receptor Inverse Agonist Site," *J. Med. Chem.* **33,** 2343–2357 (1990).
12. H. J. Breslin, M. J. Kukla, D. W. Ludovici, R. Mohrbacher, W. Ho, M. Miranda, J. D. Rodgers, T. K. Hitchens, G. Leo, D. A. Gauthier, C. Y. Ho, M. K. Scott, E. De Clercq, R. Pauwels, K. Andries, M. A. C. Janssen, and P. A. J. Janssen, "Synthesis and Anti-HIV-1 Activity of 4,5,6,7-Tetrahydro-5-methylimidazo-[4,5,1-jk][1,4]benzodiazepin-2(1H)-one (TIBO) Derivatives. 3," *J. Med. Chem.* **38,** 771–793 (1995).
13. T. Almoy and E. Haugland, "Calibration Methods for NIRS Instruments: A Theoretical Evaluation and Comparisons by Data Splitting and Simulations," *Appl. Spectrosc.* **48,** 327–332 (1994).
14. K. Faber and B. R. Kowalski, "Propagation of Measurement Errors for the Validation of Predictions Obtained by Principal Component Regression and Partial Least Squares," *J. Chemometrics* **11,** 181–238 (1997).
15. C. J. Lovely, A. S. Bhat, H. D. Coughenour, N. C. Gilbert, and R. W. Brueggeimeier, "Synthesis and Biological Evaluation of 4-(Hydroxyalkyl)estradiols and Related Compounds," *J. Med. Chem.* **40,** 3756–3764 (1997).
16. A. Pavesi, "Relationships Between Transcriptional and Translational Control of Gene Expression in *Saccharomyces cerevisiae*: A Multiple Regression Analysis," *J. Mol. Evol.* **48,** 133–141 (1999).
17. K. A. Cole, D. B. Krizman, and M. R. Emmert-Buck, "The Genetics of Cancer—a 3D Model," *Genetics* (a *Nature* publication) **21,** 38–41 (1999).
18. E. A. Winzeler, D. R. Richards, A. R. Conway, A. L. Goldstein, S. Kalman, M. J. McCullough, J. H. McCusker, D. A. Stevens, L. Wodicka, D. J. Lockhart, and R. W. Davis, "Direct Allelic Variation Scanning of the Yeast Genome," *Science* **281,** 1194–1197 (1998).
19. A. R. Katritzky, U. Maran, V. S. Lobanov, and M. Karelson, "Structurally Diverse Quantitative Structure-Property Relationship Correlations of Technologically Relevant Physical Properties," *J. Chem. Inf. Comput. Sci.* **40,** 1–18 (2000).
20. S. Spector, M. Rosconi, and D. P. Raleigh, "Conformational Analysis of Peptide Fragments Derived from the Peripheral Subunit-Binding Domain from the Pyruvate Dehydrogenase Multienzyme Complex of Bacillus Stearothermophilus: Evidence for Nonrandom Structure in the Unfolded State," *Biopolymers* **49,** 29–40 (1999).
21. S. Wold, "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics* **20,** 397–405 (1978).
22. S. Wold, "Validation of QSAR's," *Quant. Struct. Act. Relat.* **10,** 191–193 (1991).
23. I. T. Jolliffe, *Principal Component Analysis*, Springer-Verlag, New York, 1986.
24. W. H. Press, S. A. Teukolski, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd Ed., Cambridge University Press, New York, 1992.
25. J. E. Freund, *Mathematical Statistics*, 5th Ed., Prentice-Hall, Inc., Englewood Cliffs, NJ, 1992.
26. P. R. Bevington, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill Book Co., Inc., New York, 1969.

27. B. Efron, *Jackknife, the Bootstrap & Other Resampling Plans*, Society for Industrial and Applied Mathematics, Philadelphia, 1982.
28. S. J. Press, *Bayesian Statistics: Principles, Models and Applications*, John Wiley & Sons, Inc., New York, 1989.
29. W. F. Lott, "The Optimal Set of Principal Component [30] Restrictions on a Least-Squares Regression," *Commun. Statist.* **2,** 449–464 (1973).
30. H. Wold, "Nonlinear Estimation by Iterative Least Squares Procedures," F. David, Ed., *Research Papers in Statistics*, John Wiley, New York, 1966, pp. 411–444.

Received July 24, 2000; accepted for publication October 20, 2001

Daniel E. Platt *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (watplatt@us.ibm.com).* Dr. Platt is a Research Staff Member in the Bioinformatics and Pattern Discovery group of the Computational Biology Center at the IBM Thomas J. Watson Research Center. He joined IBM in 1988 and subsequently, in 1992, obtained a Ph.D. in physics from Emory University, with thesis work in the area of condensed-matter physics. At IBM, he has worked in the Computational Biology Center since its inception. Dr. Platt's interests have ranged from deposition and diffusion-limited growth processes and scale invariance to statistical modeling, bioinformatics, and biophysics.

Laxmi Parida *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (parida@us.ibm.com).* Dr. Parida is a Research Staff Member and works with the Bioinformatics and Pattern Discovery group of the Computational Biology Center at the IBM Thomas J. Watson Research Center. She received an M.S. in 1995 and a Ph.D. in 1998, both in computer science, from the Courant Institute of Mathematical Science at New York University. She joined the IBM Research Division in 1998 and has been working on computational problems arising in biology.

Yuan Gao *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (yuangao@us.ibm.com).* Dr. Gao is a Research Associate in the Bioinformatics and Pattern Discovery group of the Computational Biology Center at the Thomas J. Watson Research Center. He received his B.S. degree from Beijing University in 1992 and attended the University of Tennessee at Memphis for graduate study. After finishing his M.S. degree in biochemistry, he carried out his Ph.D. degree studies at the Department of Mathematical Sciences of the University of Memphis, completing work on the degree at the Thomas J. Watson Research Center. His primary research interests are motif-based protein structure and function prediction.

Aris Floratos *First Genetic Trust, Inc., 9 Polito Avenue, Lyndhurst, New Jersey 07071 (afloratos@firstgenetic.net).* Dr. Floratos is Director of Bioinformatics of the First Genetic Trust Corporation, responsible for the development of next-generation tools for analyzing genetic data. While the work described in this paper was being carried out, he was a Research Staff Member in the Computational Biology Center at the IBM Thomas J. Watson Research Center in Yorktown Heights, New York. He received a B.S. degree in computer science and engineering from the University of Patras, Greece, in 1991, and M.S. and Ph.D. degrees in computer science from New York University in 1995 and 1999, respectively. Dr. Floratos joined IBM at the Thomas J. Watson Research Center in 1996. His current work focuses on the application of sophisticated algorithmic and statistical tools for the analysis of data obtained from genome studies.

Isidore Rigoutsos *IBM Research Division, Thomas J. Watson Research Center, P.O. Box 218, Yorktown Heights, New York 10598 (rigoutso@us.ibm.com).* Dr. Rigoutsos manages the Bioinformatics and Pattern Discovery group in the Computational Biology Center of the Thomas J. Watson Research Center. He received a B.S. degree in physics from the University of Athens, Greece, and M.S. and Ph.D. degrees in computer science from the Courant Institute of Mathematical Sciences at New York University. He is currently a Visiting Lecturer in the Department of Chemical Engineering at the Massachusetts Institute of Technology. He has received a Fulbright Foundation Fellowship and has held an Adjunct Professor appointment in the Computer Science Department at New York University. He is the author or coauthor of seven patents and numerous technical papers. Dr. Rigoutsos is a member of the Institute of Electrical and Electronics Engineers (IEEE), the IEEE Computer Society, and the American Association for the Advancement of Science.