The Construction and Assessment of a Statistical Model for the Prediction of Protein Assay Data

J. Pittman,*,§ J. Sacks,† and S. Stanley Young‡

Institute of Statistics and Decision Sciences, Duke University, Durham, North Carolina 27708, National Institute of Statistical Sciences, P.O. Box 14006, Research Triangle Park, North Carolina 27709, and GlaxoSmithKline, Inc., Research Triangle Park, North Carolina 27709

Received October 3, 2001

The focus of this work is the development of a statistical model for a bioinformatics database whose distinctive structure makes model assessment an interesting and challenging problem. The key components of the statistical methodology, including a fast approximation to the singular value decomposition and the use of adaptive spline modeling and tree-based methods, are described, and preliminary results are presented. These results are shown to compare favorably to selected results achieved using comparitive methods. An attempt to determine the predictive ability of the model through the use of cross-validation experiments is discussed. In conclusion a synopsis of the results of these experiments and their implications for the analysis of bioinformatic databases in general is presented.

1. INTRODUCTION

Significant scientific advances are being made in the study and understanding of the human genome. Data from genome sequencing experiments and microarray studies will accelerate the production of new protein structures. Biologists may use these data to discern the function of such structures and to identify new targets for drug development from the proteins of genes that are highly expressed in cells of interest.

Structural biologists hope to eventually learn enough about protein function to be able to speed up drug design by setting up protein structure factories.[1] Because systematic, brute-force testing of every protein is logistically and financially infeasible,[2] and many of the most interesting drug targets are not currently amenable to high-throughput approaches,[3] more incisive methods are needed. This will require a fuller development of mathematical, computational and statistical techniques for predictive modeling of protein assay data in order to provide information about the interaction between drug compounds and untested proteins. The use of computational docking algorithms[2] and various statistical methods for microarray data[4,5] have already indicated the potential value of such techniques.

Given a data matrix of activity values for a set of compounds and a set of proteins (a matrix entry is the activity of a specific compound against a specific protein) the goal is to build a statistical model capable of predicting compound/protein activities for yet to be tested pairs of compounds and proteins. The use of both compound and protein structural information in making these predictions is important, leading to a number of challenges due to the complexity of protein and compound structures as well as the complexity of their interactions. The methods employed here to meet these challenges will be described within the context of a test bed example consisting of data from a set of (proprietary) drug-like compounds tested against an array of (proprietary) proteins from a family of related proteins. This example identifies issues that must be addressed in establishing a methodology transportable to other settings.

To identify the relationship between the activity data and the structural information (descriptors) about the compounds and the proteins for prediction purposes, an activity response surface must be built in the descriptor space as a model for the response. The approach taken here is to (1) reduce the dimension of descriptor spaces, (2) construct a response function using a flexible, nonparametric regression method and (3) evaluate the resulting model through cross-validation[9] and out-of-sample predictions. Each of these elements is discussed in detail. The test bed example is described in Section 2, and the problem of dimension reduction and how it is addressed are discussed in Section 3. Section 4 provides an overview of the flexible modeling methods employed and their use in the test bed. Section 5 is focused on the validation of the constructed statistical model and comparison of its performance with those of other methods. Section 6 highlights a discussion of the findings and directions for future research.

2. TEST BED

For each of a collection of $n_c = 576$ drug-like compounds a binding assay was made with each of $n_q = 10$ proteins. The binding response (activity) is a measure of affinity between a compound (or ligand) and a protein (or receptor), namely, the $-\log_{10}$ concentration of a compound that inhibits the response or binding of a submaximal concentration of a competing compound to a protein by 50% ($p\mathrm{IC}_{50}$). A large binding response or activity value indicates that the compound of interest is able to competitively displace a standard compound from its binding site, thereby counteracting its effect. The test bed response variable is the binding response arranged as a compound × protein matrix. The proteins and compounds were selected for their potential use in pharma-

* Corresponding author phone: (919)684-4447; e-mail: jennifer@stat.Duke.edu.
† National Institute of Statistical Sciences.
‡ GlaxoSmithKline, Inc.
§ Duke University.

10.1021/ci0103828 CCC: $22.00
Published on Web 05/28/2002

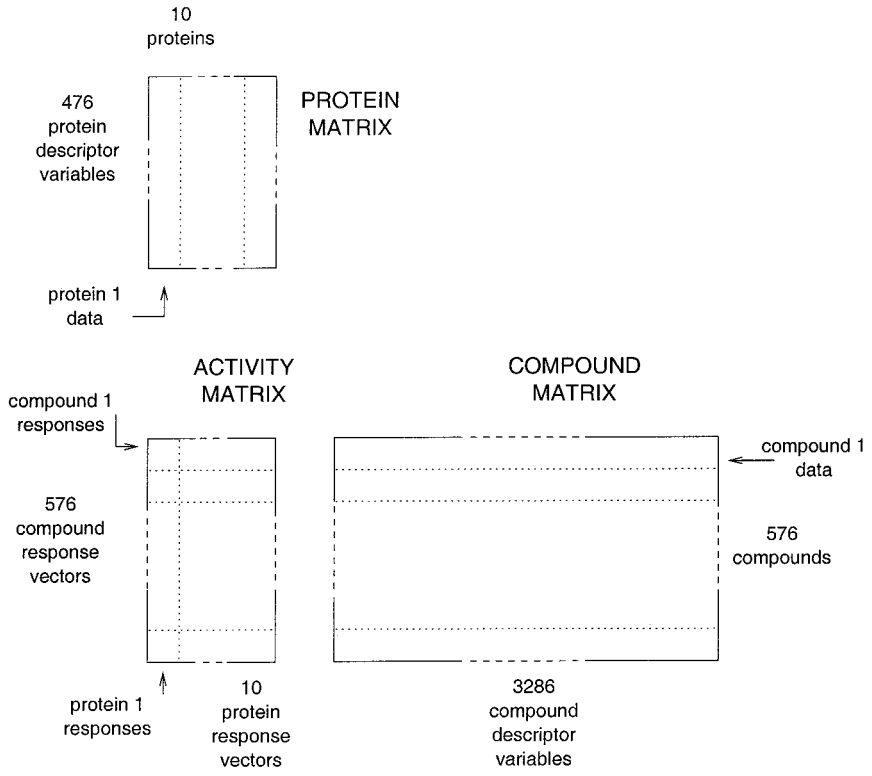

Figure 1. Three-way data structure of test bed data set.

ceutical development. For various reasons, the available 576 × 10 matrix of data contained missing values; how these "missing" observations were handled is discussed below in Section 4.1 where further specifics of the data are also described.

Besides the response matrix the test bed contains both a compound and a protein descriptor matrix. A set of compound descriptors that has been effective in the analysis of high throughput screening data[12] identifies the presence/absence of topological features in each compound. A total of $p_c = 3286$ such identifiers were used to describe the 576 compounds; hence the compound descriptor matrix is a 576 × 3286 (compounds × descriptors) matrix of 1's and 0's; a 1 in the (i, j) entry of the matrix indicates the presence of atom pair (or other structural feature) j in compound i.[13] Unfortunately the science of protein description is rather complex and as of yet there is no standard technique or method for description. This, combined with the considerable expense of selecting a protein target for study and developing an assay, has made the search for effective descriptors and proteins useful in pharmaceutical applications highly competitive. As a result the specific proteins and protein descriptors in the test bed are proprietary. The number of protein descriptors is $p_q = 476$ and hence the protein descriptor matrix is 10 × 476 (proteins × descriptors) with real number entries. A schematic of the test bed data set is shown in Figure 1.

This data structure is described as "three-way" as it contains three data matrices.

3. DIMENSION REDUCTION

The descriptor data live in high dimensional spaces—spaces which, certainly for compound sets in other examples, can have tens of thousands of descriptors and hundreds of thousands of compounds. Protein descriptor matrices, due to the costs of data collection, usually follow the "large p, small n" paradigm where the number of columns (the descriptors), p, far exceeds the number of observations or rows, n. As p increases a data set of fixed size n becomes increasingly more sparse—the so-called "curse of dimensionality"[14]—making prediction in the descriptor space more difficult. Projecting the data into a lower dimensional space, e.g., dimension reduction via singular value decomposition (SVD),[6,15] can be useful, if not essential, for making these data sets amenable to statistical analyses.

Let our descriptor matrix $\mathbf{X}$ have n rows and p columns where $n < < p$. Then $\mathbf{X}$ is rank deficient, i.e., the dimension of the descriptor space is r where $r < \min(n, p)$. By matrix theory,[6,15]

$$\mathbf{X} = \mathbf{U}\Lambda\mathbf{V}^{\mathbf{T}}$$

where $\mathbf{U}$ is a $n \times r$ column-orthonormal matrix, Λ is a $r \times r$ diagonal matrix and $\mathbf{V}$ is a $p \times r$ column-orthonormal matrix. The matrices $\mathbf{U}$, Λ, and $\mathbf{V}$ are components of the SVD of $\mathbf{X}$. The information contained in the original predictor variables has been recoded in a much lower-dimensional space of a new set of predictor variables, the columns of $(\mathbf{U}\Lambda)$. These columns are referred to as the principal components (PCs)[16] of $\mathbf{X}$.

In PC analysis the matrix $\mathbf{U}$ is called the factor matrix of $\mathbf{X}$, the diagonal elements of Λ are the singular values of $\mathbf{X}$ listed in decreasing order ($\lambda_{11} \geq \lambda_{22} \geq \cdots \geq \lambda_{rr} \geq 0$) and $\mathbf{V}$ is called the loadings matrix. Each of the r PCs is essentially a linear combination of the original p predictor variables, where the coefficients of each linear combination are given in the r columns of $\mathbf{V}$ or the r loading vectors.[17] Hence SVD is a tool for dimension reduction that can indicate (1) the relative importance of the original variables in the construc-

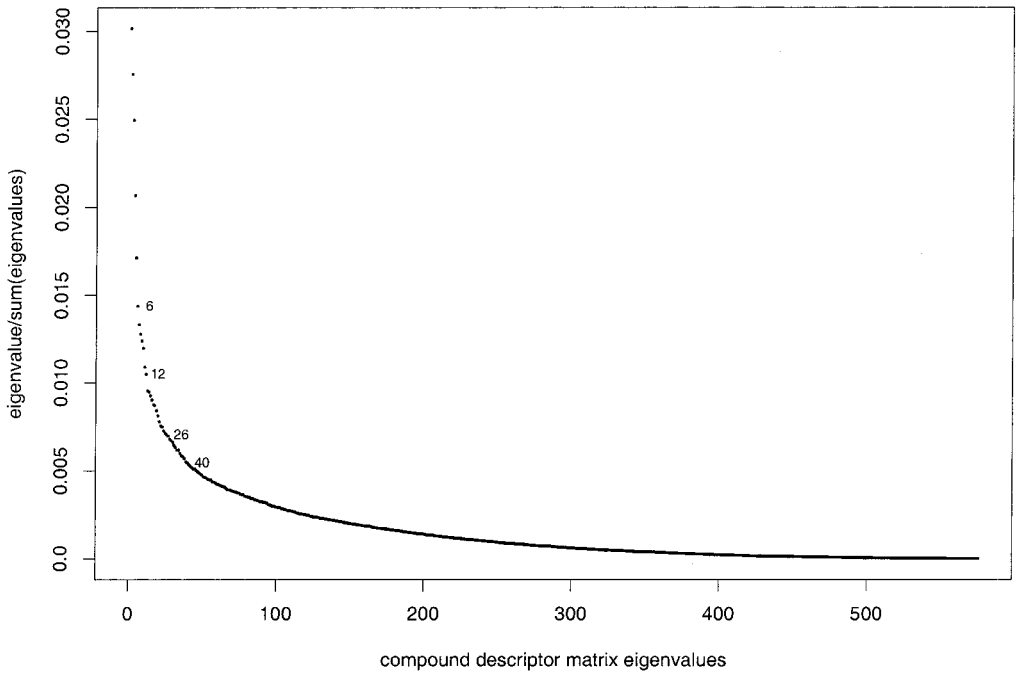

Figure 2. Eigenvalues of the compound descriptor matrix of the test bed data set.

tion of a smaller set of vectors that span the design space and (2) patterns and relationships among the original variables through the values in the loading matrix.

The SVDs of the test bed descriptor matrices yield 576 compound PCs and 10 protein PCs. A scree plot,[18] a plot of λ_{ii}^2 or $\lambda_{ii}^2/\sum_{i=1}^{r} \lambda_{ii}^2$ versus i, is used to decide how many components to retain by revealing a break between the "larger" and "smaller" eigenvalues represented by a sharp bend in the curve. The PCs corresponding to the eigenvalues appearing before the curve are retained; the remaining PCs are considered "noise". Figure 2 is a scree plot of the eigenvalues of the compound descriptor matrix and, although the bend is not sharp, a good estimate can be made of the number of significant components. After examining scree plots of the compound and protein eigenvalues 12 compound PCs and five protein PCs corresponding to the largest singular values were selected for inclusion in a model of the test bed.

A schematic diagram of the original data matrices and the two matrices of selected principal components is depicted in Figure 3. The matrices enclosed in circles are the components of the data set after dimension reduction.

The 17 selected PCs and their interactions are used as candidate variables in constructing the predictive regression model of Section 4.1. Unfortunately the SVD, albeit a valuable analytical and computational tool, cannot take advantage of matrix properties such as sparseness to reduce computational and storage demands and is difficult to update/downdate when observations are added/removed from the data set (as in cross-validation).[19] This makes SVD computationally expensive and impractical for large data sets. Decompositions that provide reliable estimates of rank and relevant subspaces and are considerably more efficient relative to SVD are rank-revealing two-sided orthogonal decompositions, also referred to as complete orthogonal decompositions[6] or UTV decompositions.[15,19] The UTV decomposition and its use in the cross-validation studies of the compounds in our test bed will be discussed in further detail in Section 5.2.

4. ADAPTIVE NONPARAMETRIC REGRESSION

In the statistical application of interest a modeling technique is needed which can capture a relationship between a response variable Y and a set of one or more predictor variables $X_1, ..., X_p$ that is more complex than a simple linear relationship. There are numerous statistical modeling techniques available for this purpose; see, for example, refs 20−22. Nonparametric regression is one such technique which has been successful in characterizing features of data sets that could not be described by other means. Among the numerous modeling techniques which fit the nonparametric regression paradigm, those which use piecewise functions of the predictors as model terms can generate models in moderately high dimensions which are flexible yet in which model terms are represented in strictly low dimensions.[20]

One well-known nonparametric regression algorithm is Multivariate Adaptive Regression Splines [MARS][8] in which predictor variables are represented by piecewise linear (or cubic) functions and a subset of these predictors and their interactions are fit to the data as terms in a regression model in several dimensions. A one-dimensional term is interpreted as a main effect and a two-dimensional term is interpreted as an interaction between its component main effects. This model can be used like any other regression model for data analysis and prediction. More details about the MARS algorithm and its scalability are provided in Appendix A.

4.1. Three-Way Data. For the statistical analysis the 12 compound PCs (ordered by decreasing singular values) were labeled as variables #1−12 and the five protein PCs (also ordered by decreasing singular values) were labeled as variables #13−17. To involve both the compound and protein PC descriptors, the three-way data structure was reexpressed in a two-way format using matrix manipulations. Let a_{ij} represent the (i, j)th element of the activity matrix (the

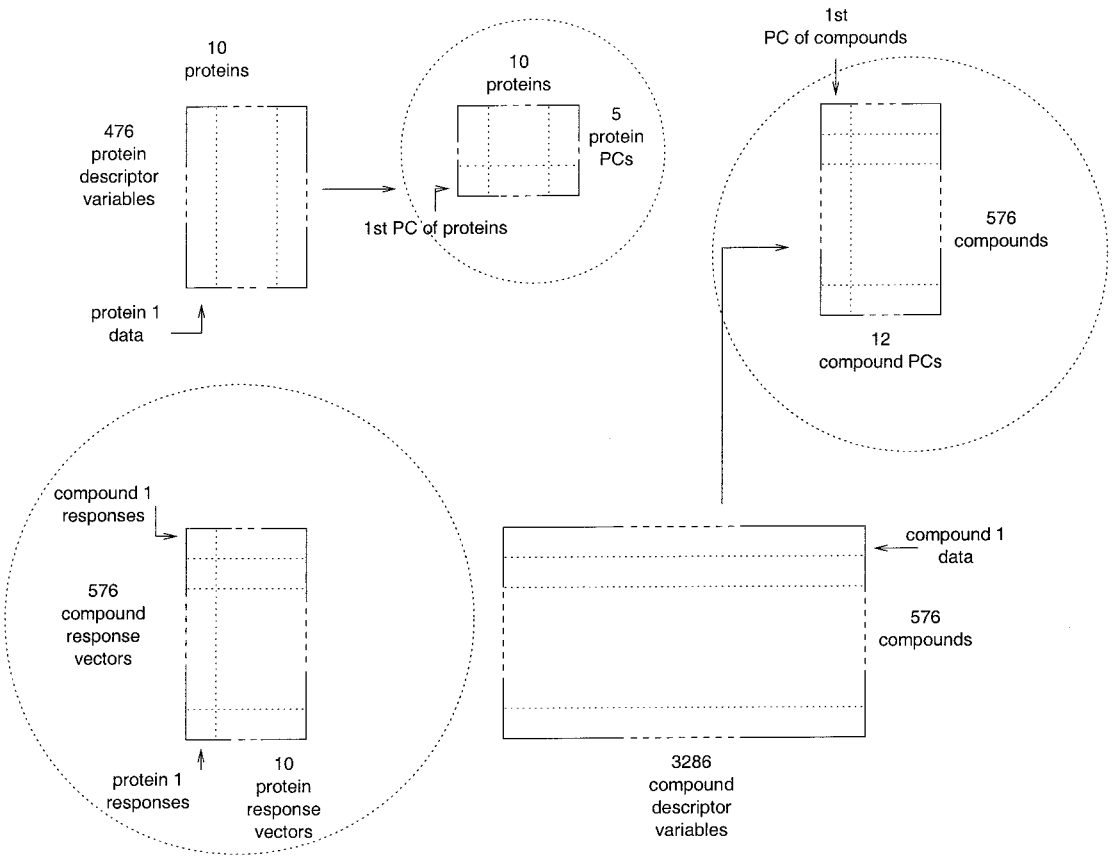

Figure 3. Three-way data structure and results of dimension reduction.

response of the *i*th compound and the *j*th protein), c_{ij} represent the (*i*, *j*)th element of the compound PC matrix (the *i*th element of the *j*th compound PC), and q_{ij} represent the (*i*, *j*)th element of the protein PC matrix (the *i*th element of the *j*th protein PC). The response and descriptor matrices after reexpression may be depicted as seen below ($n_c = 576$, $n_q = 10$, $t = 12$).

$$\begin{bmatrix} a_{11} \\ a_{12} \\ \vdots \\ a_1 n_q \\ a_{11} \\ \vdots \\ a_{1_{nc}} 1 \\ \vdots \\ a_{11} \end{bmatrix} \begin{bmatrix} c_{11} & c_{12} & \cdots & c_1 t & q_{11} & q_{11} & \cdots & q_{15} \\ c_{11} & c_{12} & \cdots & c_{1t} & q_{21} & q_{22} & \cdots & q_{25} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ c_{11} & c_{12} & \cdots & c_{1t} & q_{n_q 1} & q_{n_q 2} & \cdots & q_{n_q 5} \\ c_{21} & c_{22} & \cdots & c_{2t} & q_{11} & q_{12} & \cdots & q_{15} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ c_{n_c 1} & c_{n_c 2} & \cdots & c_{n_c t} & q_{11} & q_{12} & \cdots & q_{15} \\ \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ c_{n_c 1} & c_{n_c 2} & \cdots & c_{n_c t} & q_{n_q 1} & q_{n_q 2} & \cdots & q_{n_q 5} \end{bmatrix}$$

Each cell of the original response matrix corresponds to a row of the new descriptor matrix; the (*i*, *j*)th element of the combined descriptor matrix contains the $(\overline{i/n_q})$th row of the original compound PC matrix followed by the *k*th row of the original protein PC matrix, where $k = mod\,(j, n_q)$ if $mod\,(j, n_q) \neq 0$ and $k = n_q$ otherwise. The new response vector has 5760 entries and the new PC descriptor matrix has 5760 rows and 17 columns.

In this two-way format missing or erroneous activity values can be easily removed by row deletion. This capability is extremely useful as it allows the use of incomplete protein/compound profile data. In the test bed data set 1149 data values were missing or incomplete, leaving $N = 4611$ observations for analysis. This final data set contains ($\leq p_q$) observations or rows for each compound or, similarly, ($\leq p_c$) observations for each protein. The two-way format of our test bed resembles that of the data set shown in Figure 4 which contains activity values for only 8 of 10 proteins for compound 1 and activity values for only 574 of 576 compounds for protein 10.

4.2. Test Bed Model. MARS version 3.6[8,23] was applied to the test bed, allowing for main effects and two-way interactions and with all parameters set at their default values (see Appendix 6.1). The results for the entire test bed are displayed in Figure 5, and the fitted values appear to be quite good with $R^2 = 0.64$. The PC descriptors found to be of most importance in terms of influence on model fit, as percent relative to the most important descriptor, are protein PC #15 (100%), compound PC #1 (69%), protein PC #17 (37%), compound PC #2 (30%), protein PC #14 (28%) and compound PC #6 (26%). Of the 10 most important model terms 7 are interaction terms of which 6 are interactions between a compound PC and a protein PC (see Figure 6). A significant compound/protein interaction term indicates that the bioassay response depends on specific compound features in combination with specific protein features; these features may be identified by using the PC loading vectors to map descriptors to features. It was noticed that the compound PC #1 has heavier loadings on descriptors that are characteristic of compounds active against proteins #2 and #3—the two proteins with the most active responses. Similarly compound PC #2 has heavier loadings on descriptors that are charac-

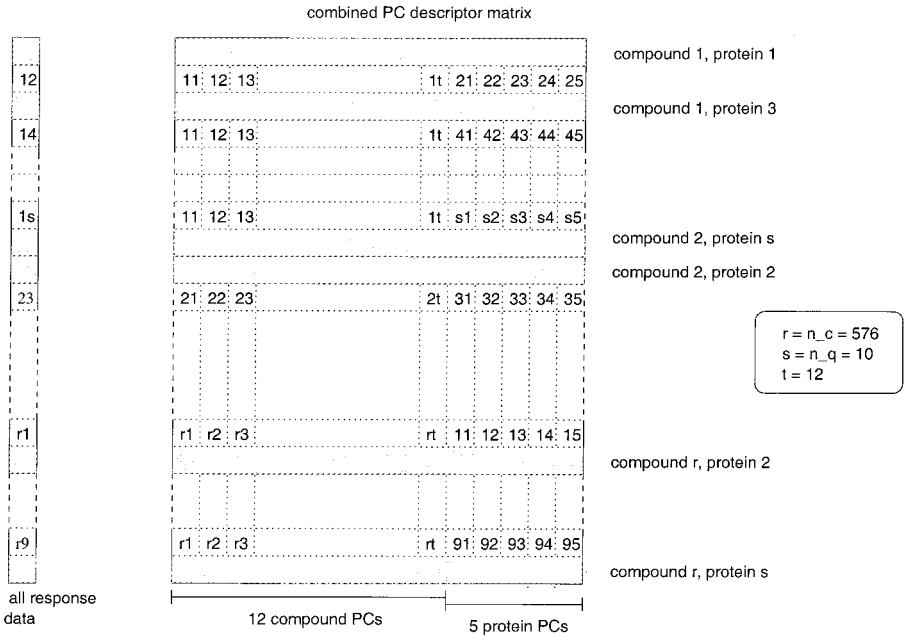

Figure 4. Removal of missing or incomplete data from two-way data structure.

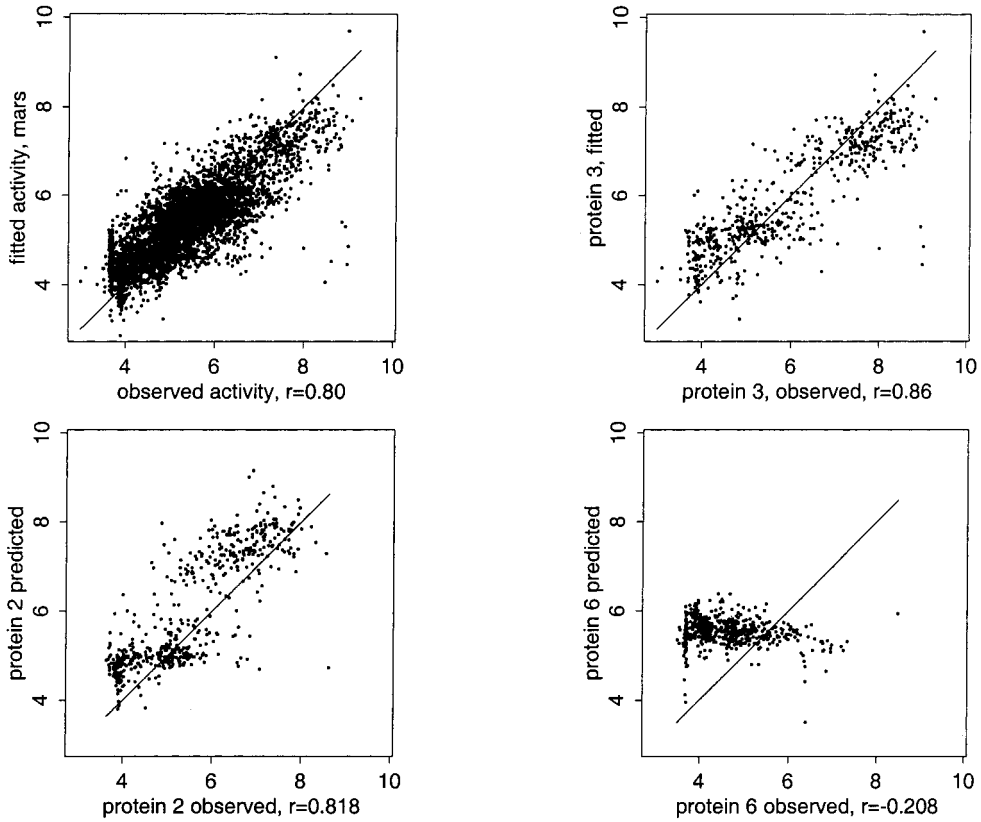

Figure 5. Observed and predicted values from test bed MARS model for noncv (top row) and cv (bottom row) studies. Reference line of 45 degrees is added.

teristic of compounds active against protein #7—a protein which is relatively uncorrelated with the other proteins (highest correlation = $r = 0.496$). Results such as these can help identify compound descriptors whose presence/absence influences how a compound behaves in the presence of various proteins. Interpretation of PC loading vectors in principle is straightforward, as high loadings imply greater influence, although in practice interpretation can be problematic when there are numerous variables or lack of subject information. The interpretation of the protein PC loading vectors was much more difficult, e.g., the third protein PC (PC #15) has many large loadings and could not be interpreted given almost no subject knowledge.

The variability in average activity level among assays can be quite high so the fitted response values for each protein assay were examined separately. It was expected that if the model was performing well the quality of the fitted values would be somewhat higher for proteins with a wider range of activity levels, higher correlations with other proteins or smaller numbers of missing or incomplete observations. The

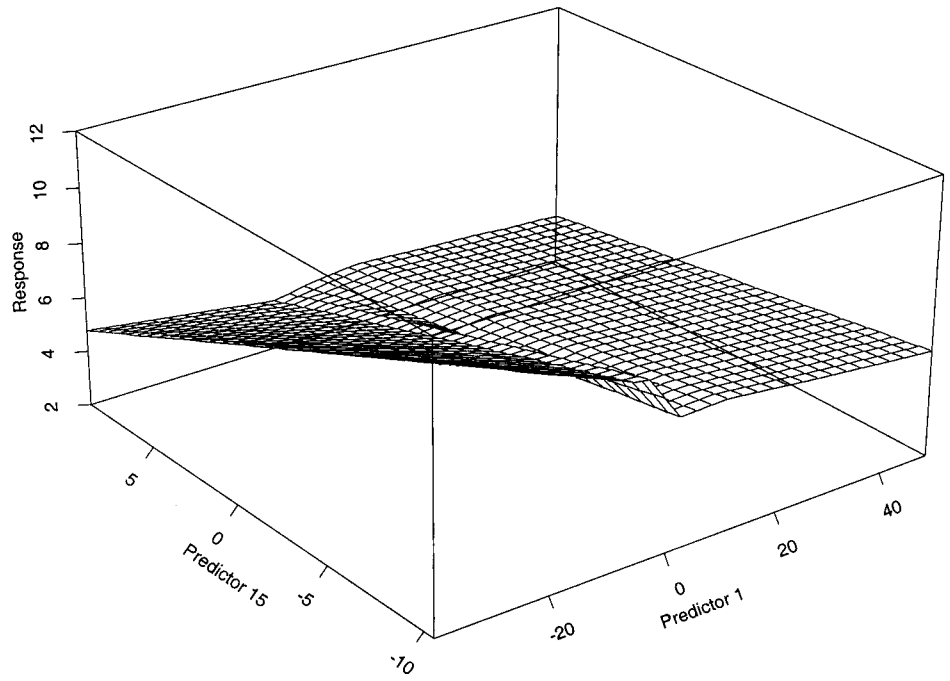

Figure 6. Plot of significant interaction between PC variables #1 (compound) and #15 (protein).

Table 1. Test Bed Protein Results

	protein results									
protein	1	2	3	4	5	6	7	8	9	10
nobs	553	536	547	362	390	557	503	339	548	276
max(cor)	0.528	0.903	0.903	0.622	0.652	0.210	0.448	0.320	0.574	0.665
res(cor)	0.479	0.838	0.864	0.202	0.370	0.709	0.347	0.293	0.701	0.331

amount of improvement was substantial as seen in Table 1: for example, the fitted values for protein 3 which has 29 incomplete observations, a maximum correlation with another protein of $r = 0.903$ (max(cor)) and an activity value range of [3, 9.26] (see Figure 5) are of a much higher quality than those for protein 5 which has 186 incomplete observations, a maximum correlation with another protein of $r = 0.652$ (max(cor)) and an activity value range of [4.75, 6.08] (res(cor) = correlation(observed,predicted)). The only noticeable trend in the compound results was expected given a regression method: a tendency to slightly overestimate those compounds with the lowest average activity values and slightly underestimate those compounds with the highest average activity values (results not shown).

5. MODEL VALIDATION

Modern regression techniques have the flexibility to exploit information in the data inaccessible to more traditional methods. This flexibility, however, can lead to overfitting and vulnerability to the influence of high leverage or extreme data points.[24] As a result the model may be misleading and unfit for predicting new observations. Cross-validation and out-of-sample predictions can be used to provide a measure of the quality of the fitted model by providing predicted values for the original data set are created that can give an honest assessment of the predictive power of a model.[4]

Assuming that the original data represent a random sample leave-one-out (loo) cross-validation[25] yields predictive values that are unbiased, i.e.,

$$E[\hat{y}_i] = E[y_i]$$

where $E[y_i]$ is based on the full sample. However, loo often tends toward overfitting[16] leading to models that contain too many factors or with overly optimistic prediction errors. A useful alternative is $\boldsymbol{k}$-fold cross-validation[16,26,27]; this is implemented with the test bed by holding out sets of compounds (sets of rows in the compound × protein activity data matrix). Because there are (today) few proteins in a typical study, only single proteins (single columns) are held out. It is noted that assays are often made with compounds that are chemical "twins" or "tuplets" as well as with compounds that are structurally unrelated; the effect of twins in cross-validation is considered by applying a hierarchical clustering technique referred to as MONA (monothetic analysis)[29] to the compound descriptor matrix to classify the compounds into distinct structural groups for prediction (MONA is described in more detail in Section 5.2).

Cross-validation analyses are performed using MARS, *k*-nearest neighbor (*k*-NN), and main effect models with linear and cubic spline terms (via functions knn and lm from the program R[28]) for comparison. To confirm the MARS results, in Section 5.3 a data set containing assay values for a new set of compounds is used as a model test set while the original data set is used as a training set for developing the model. The model test set contains activity values for 249 compounds, each of which has been assayed against a subset of the test bed protein targets.

5.1. Prediction of Protein Activity. For each of the 10 test bed proteins that are left out, the SVD of the remaining 476×9 protein descriptor matrix is computed. All compounds are retained so there is but one SVD to compute for the full

Table 2. Cross-Validation Protein Results (Observed,Predicted)

	cross-validation protein results									
protein	1	2	3	4	5	6	7	8	9	10
mars	0.354	0.818	0.663	0.150	0.065	-0.208	0.322	0.261	0.142	0.297
lm	0.427	0.733	0.184	0.08	-0.192	-0.250	0.010	0.195	0.565	0.241
k-NN	-0.06	0.101	0.066	0.117	-0.290	-0.053	0.103	0.085	0.095	-0.036
k	5	5	1	5	1	10	10	5	10	5

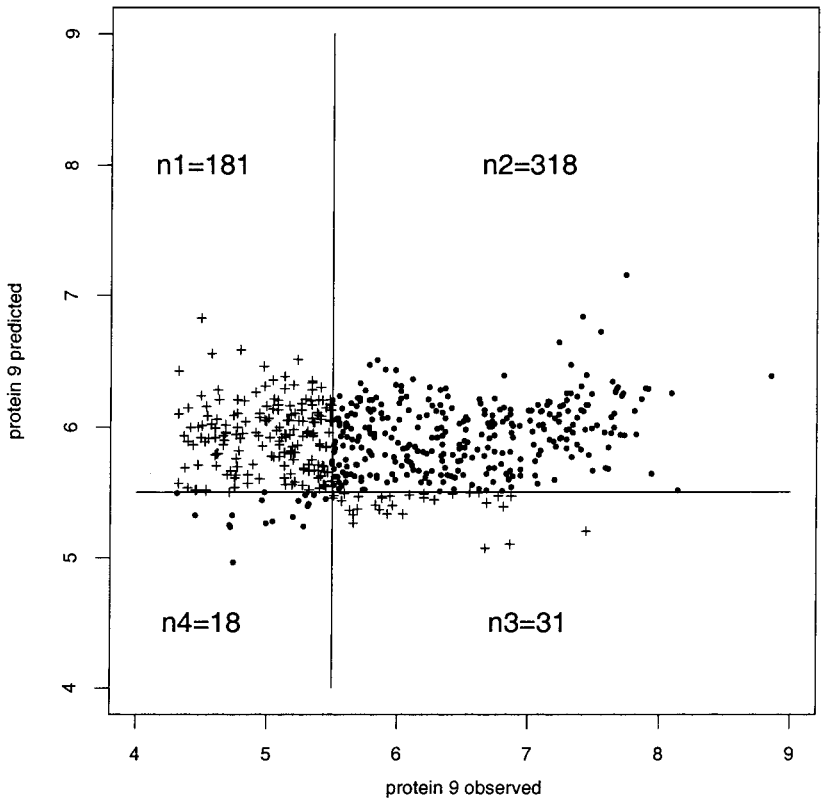

Figure 7. Cross-validated predictions for protein 9 classified as match or nonmatch.

compound descriptor matrix. The cross-validation results show that due to the variability among proteins the predictive power of the model depends substantially on the specific protein in question (see Figure 5 and Table 2). For example, the highest correlation that the second protein has with another protein in the set is 0.903. As expected, a scatterplot of predicted vs observed shows good predictive capability ($r = 0.818$). The sixth protein, in contrast, has a highest correlation value of 0.210, and the predicted responses are quite poor ($r = -0.208$). The model is capable of using information about a protein in generating predictions, when such information exists, without introducing structure where little or no structural information exists.

With only moderate correlation between training and test proteins the predicted activity values are active (> 5.5) if the corresponding observed values are active and visa versa. The observed values for protein 9, for example, have a highest correlation value with another protein of $r = 0.574$; yet as seen in Figure 7 the predicted values match the active/nonactive levels of the observed values for 61% of the data points (and for 91% of the "active" observed values). By comparison, for protein 2 the active/nonactive levels match for 86% of the data points and for 88% of the active observed values (results not shown).

In response to the suggestion that simpler methods could achieve similar cross-validation results, both *k*-NN ($k =$ 1,3,5,10) and main effect models with linear and cubic spline terms were used in the same protein cross-validation studies. The best cross-validation results achieved for all proteins and for all three methods are shown in Table 2. In some cases the results of the competing methods rival those of MARS, but overall these results and the results shown in Figure 8 support the use of a more complicated modeling technique for prediction.

5.2. Cross-Validation Studies: Prediction of Compound Activity. Leave-out/test sets representing 10% and 20% of the compounds were selected both systematically and at random. The results for 20% sets were similar to those for 10% sets, and results for systematically selected sets were similar to those for randomly selected sets; the results for 10% sets selected systematically are discussed here. Each test group was selected to have approximately the same average response value by ordering the compounds by decreasing average activity value and selecting every tenth compound from the resulting list (starting from the *k*th compound to form the *k*th group for $k = 1, ..., 10$). In other words, the *k*th test group contained compounds (k,$10 + k$,$20 + k$,...,$560 + k$); the final six compounds were appended to the last group resulting in a total of nine groups of 57 compounds and one group of 63 compounds. The cross-validation results were quite good and consistent across groups; a prototypical set of results is shown in Figure 9.

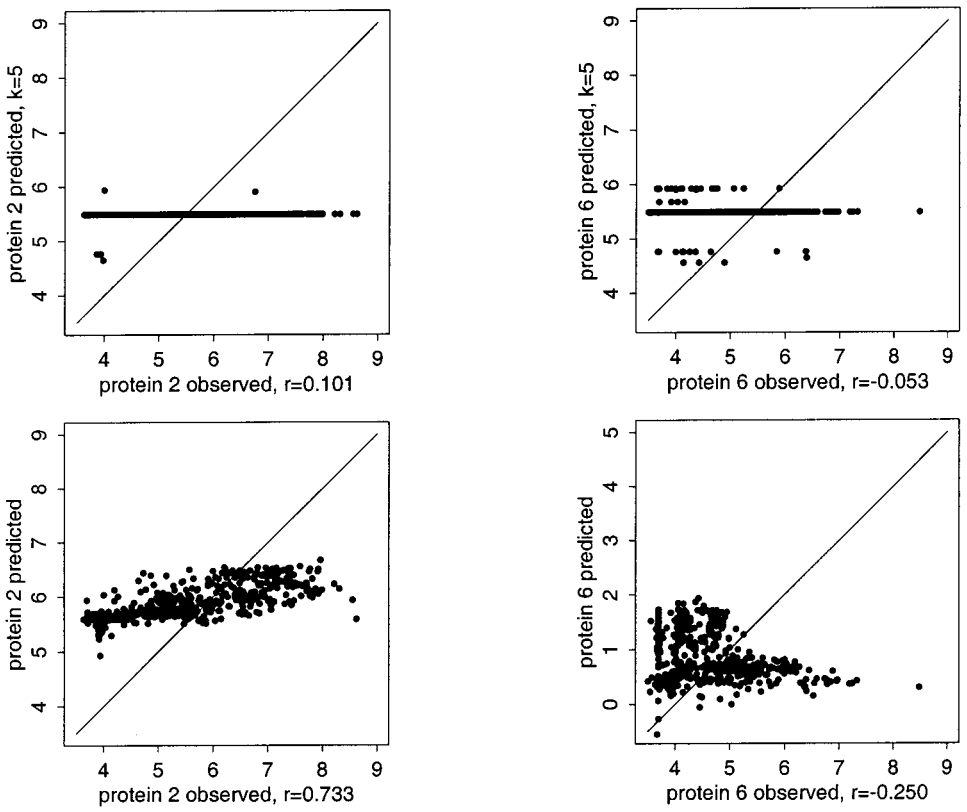

Figure 8. Cross-validated predictions for proteins 2 and 6 using *k*-NN and main effect additive models. Reference line of 45 degrees is added.

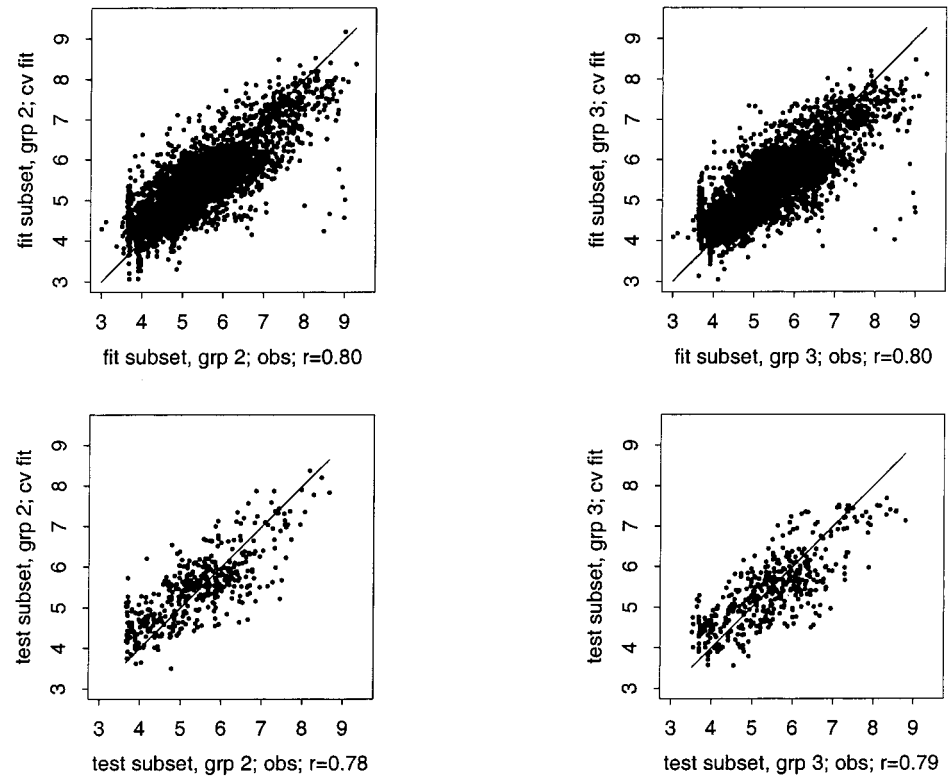

Figure 9. Fitted and observed activities from cross-validation studies of two compound groups selected systematically and containing 10% of the compounds. Reference line of 45 degrees is added.

In contrast, for the same two prototypical compound groups the best main effect linear/spline model yielded $r = 0.614$ and $r = 0.611$ for the training sets and $r = 0.301$ and $r = 0.208$ for the test sets (*k*-NN results were not competitive).

Since these results might be attributed to compounds with near exact structural similarities in the training and test sets, MONA,[29] as mentioned above, was applied to the compound descriptor matrix. MONA starts with all of the observations in a single cluster and then constructs a hierarchy of clusters

Table 3. Cross-Validation MONA Results [Cor(Observed,Predicted)]

	cross-validation MONA results											
group	1			2			3			4		
CV %	100	50	0	100	50	0	100	50	0	100	50	0
mars	0.46	0.53	0.62	0.76	0.88	0.89	0.50	0.36	0.51	0.66	0.74	0.78
k-NN	0.06	0.12	-0.07	-0.16	-0.06	-0.07	0.09	0.39	0.06	0.08	0.32	-0.11
lm	0.28	0.00	0.37	0.62	0.28	0.71	0.18	0.42	0.22	0.61	0.38	0.66

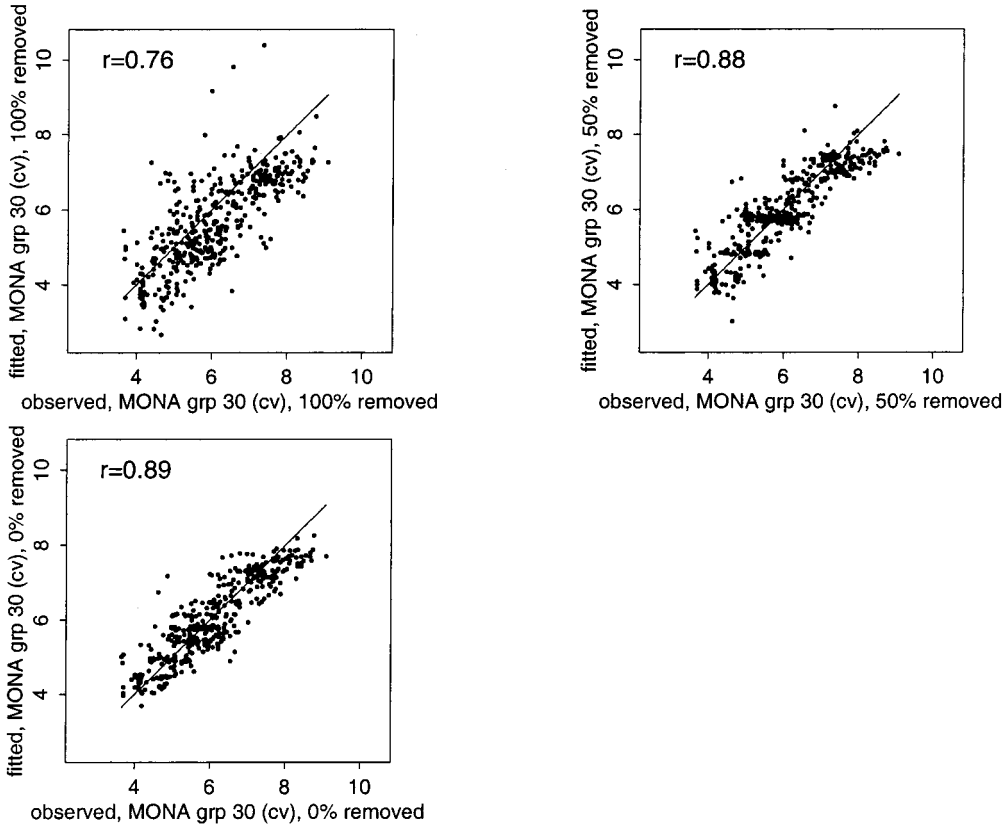

Figure 10. Fitted and observed activities from cross-validation study with MONA group of high average activity. Reference line of 45 degrees is added.

where at each stage a split of a cluster is determined by the variable with the maximal total association with all of the other variables (among the observations in the cluster being split). The association between two variables z_i and z_j is defined as $a(z_i, z_j)\times d\ (z_i, z_j) - b(z_i, z_j)\times c(z_i, z_j)$ where $a(z_i, z_j)$, $b(z_i, z_j)$, $c(z_i, z_j)$, and $d\ (z_i, z_j)$ are the entries in the contingency table of z_i and z_j: (a (d) is the number of observations where both z_i and z_j are 0 (1) and b (c) is the number of observations where z_i is 0 (1) and z_j is 1 (0). The total association for a variable z_i is the sum of its associations with all of the other variables. For our test bed the maximum number of clusters or nodes was set at 30 with a minimum node size of five compounds based on subject matter knowledge. The resulting MONA tree had 30 nodes with an average node size of 19 compounds. Of these 30 nodal sets 10 were randomly selected for cross-validation testing.

For each of these 10 selected groups three sets of predicted values were compared: predicted values when all (full cross-validation), 50% (partial cross-validation), or none (no cross-validation) of the observations in the group had been used as the test set. Summary results for these three sets of predicted values for a random subset of four of the 10 selected groups are shown in Table 3 (*k*-NN gave best results when $k = 10$). The averages and standard deviations of the correlations between mars fitted and observed values across all test groups, for the three sets of predicted values, were (0.50, 0.62, 0.67) and (0.24, 0.16, 0.14), respectively. There is a clear trend in the MARS results of correlation increasing and correlation uncertainty decreasing as fewer MONA observations are removed from the training set. The results for one MONA group are shown in Figure 10. No relationship was observed between the average activity value of a MONA group and the effect of removing observations. Main effect linear/spline model results for the MONA groups fell between those using MARS and *k*-NN methods, with comparable standard deviations and average correlations across all test groups for the three sets of predicted values of (0.44, 0.48, 0.53).

One conclusion from these results is that chemical twins in the compound set are giving a boost to the predictive power of the model. It is common knowledge in pharmaceutical laboratories that a very small percentage of all compounds are active so for each active compound discovered a number of twins (or tuplets) are created—groups of one of more compounds whose structures vary only slightly from that of the originating active compound. In assessing the predictive ability of a statistical model, the division of twins among training and test sets or the use of all twins in

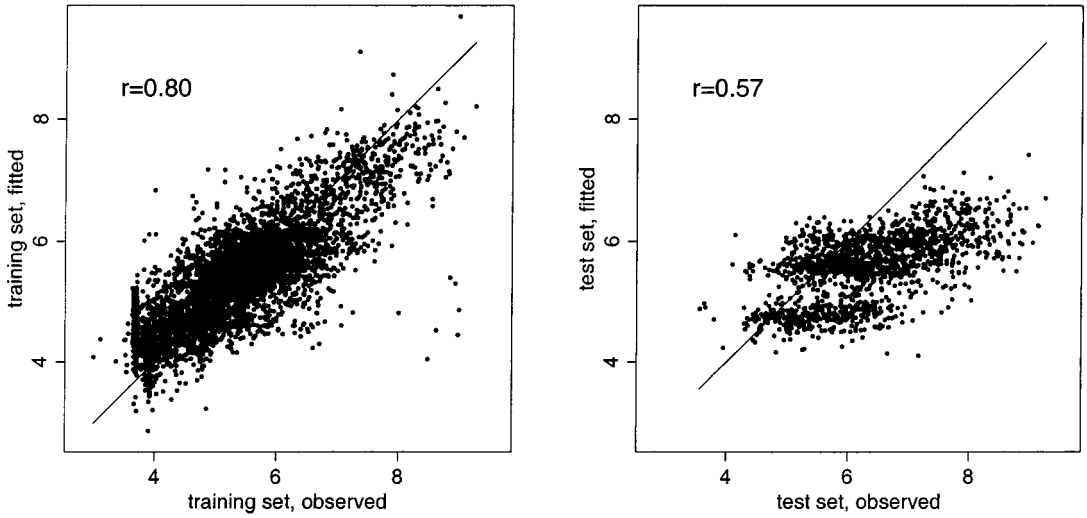

Figure 11. Observed and predicted activity values for the test bed or training data and the new or test data provided by the test bed MARS model. Reference line of 45 degrees is added.

Table 4. Cross-Validation Protein Results for New Data

protein	1	2	3	4	5	6	7	8	9	10
nobs	199	249	249	129	116	29	249	94	249	89
res(cor)	0.112	0.551	0.640	0.009	-0.104	0.335	0.028	-0.274	0.442	-0.213

either subset will bias the results. A similar property has been observed in studies of human cancer cell lines—when a hierarchical clustering algorithm is applied to the data vectors of the different cell lines, the cell lines that originated from common tissues form almost disjoint groups of closely related observations.[30] The concept of what constitutes a "fair" assessment of model validity given such a data set is unclear.

5.3. Model Validation Study: Prediction of Activity of New Compounds. As a final study the MARS model of the test bed data set (see Section 4.1) was used to predict the activity of a new set of 249 compounds against at least four of the 10 test bed proteins. These compounds were selected from a much larger set of ~26 000 compounds, the same set from which the test bed compounds were selected. ~9500 of these ~26 000 compounds were selected at random and listed by MONA class; every other set of 500 compounds was chosen from the resulting list. This was designed to provide a subset of new compounds representative of this entire set. The ~500 compounds from this subset with assay values for at least four of the 10 test bed proteins were retained, their average activity values were calculated, and the 249 compounds with the largest average activity values (a total of $N = 1651$ activity values) were chosen as the new compound set.

The 3286 binary compound descriptor variables used in the test bed (see Section 2) were obtained for the new compounds and used in calculating compound PC descriptor values for the new compounds. These PCs and the test bed protein PCs were used to form a complete 249×17 matrix of descriptor PC variable values on which the test bed model was applied. The resulting correlation between the observed and predicted values is 0.57 which is comparable with the results of the MONA cross-validation studies (see Figure 11).

As with the test bed data set the value of the predictions varied significantly with the specific protein (see Table 4).

As was found in the protein cross-validation test bed studies, the model did well in predicting whether an observed activity level was active (<5.5) or nonactive (results not shown). All of the results concerning the new compound set support the results of the test bed cross-validation studies.

6. DISCUSSION

A strategy for the analysis and prediction of biological assay data to address problems in drug discovery and molecular pharmacology has been outlined. As an example of how this methodology may be implemented, the analysis of a protein assay data set was examined in which the main goal was to construct a statistical model that could provide a prediction of the response of a novel protein to a given set of compounds or visa versa.

It is well understood that computational models are being increasingly used to provide predictions of observations of interest, and confidence in these predictions should come from comparisons with existing data.[31] Cross-validation studies as well as studies involving new data were performed which provided unbiased information about the relationship between the observed and predicted activity data. Overall the model performed better than both *k*-NN and main effect additive models. The existence of "training" compounds that belonged to the same structural group as the "test" compounds leaves open the question of how to assess the predictive capabilities of the model in a manner that will be useful and informative to interested scientists.

Future Research. Although the analysis strategy that has been outlined may be used in modeling applications outside of bioinformatics, the research focus is to develop this methodology for use in proteomics and genomics, i.e., with databases involving observations from protein assays and/or DNA microarrays. With this in mind, the choice of statistical model and fitting method as well as the selection of training and cross-validation samples deserve further

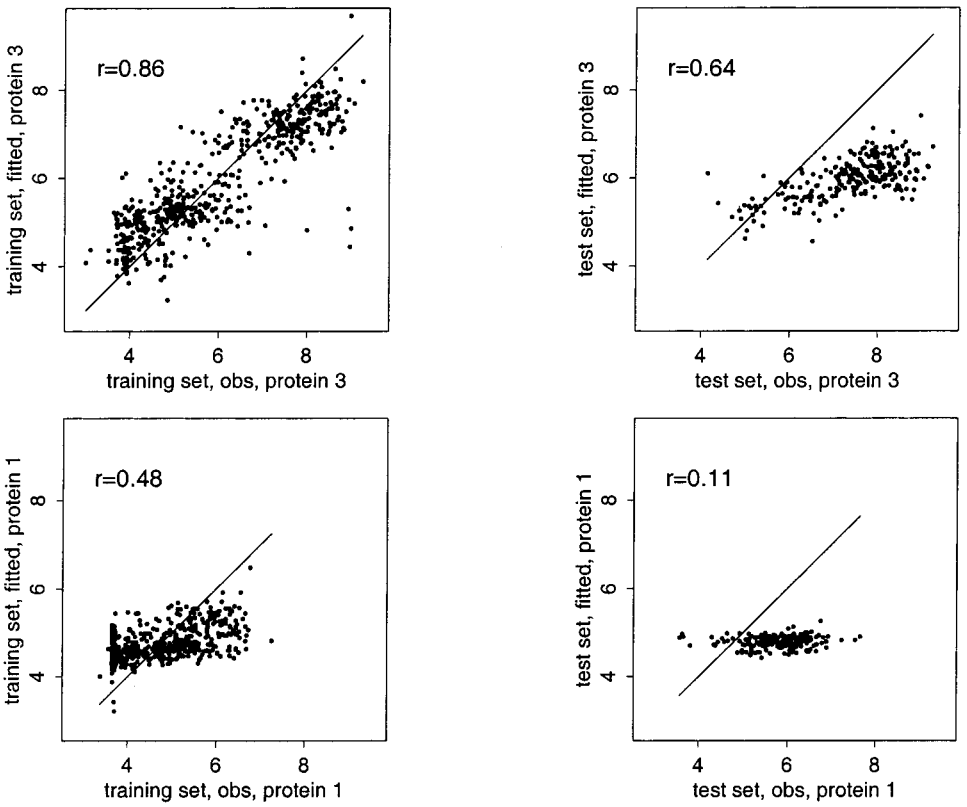

Figure 12. Observed and predicted activity values for the test bed and the new compound data for proteins 3 and 1 provided by the test bed MARS model. Reference line of 45 degrees is added.

attention. It is possible that modeling techniques not considered here, such as the use of wavelets and other clustering methods, may be appropriate for databases of this structure. It would also be constructive to develop an outline for incorporating the objectives and utilities of the users of a database into the evaluation of model predictions.

ACKNOWLEDGMENT

I would like to thank Professor Jesse Barlow of the Department of Computer Science and Engineering at The Pennsylvania State University for his helpful suggestions and insights regarding reduced-rank matrix decompositions. The research of J. Pittman was funded in part by the National Science Foundation under research Grant NSF-DMS #9700867 and in part by GlaxoSmithKline, Inc., under research contract # LGC13188 between GlaxoSmithKline, Inc., and the National Institute of Statistical Sciences. The research of Jerry Sacks was funded in part by the National Science Foundation under research Grant NSF-DMS #9700867.

APPENDIX A. MARS ALGORITHM

In Multivariate Adaptive Regression Splines [MARS][8] Friedman approaches predictor subset selection by using a stepwise knot addition/deletion strategy with univariate piecewise linear truncated power spline functions and their tensor products. A tensor product of k truncated power splines is of the form

$$\prod_{i=1}^{k} [s_i(x_{r(i)} - t_i)]_+^l$$

where $r(i)$ labels the predictor variables, t_i is a split point or knot on the corresponding variable, l is the order of the spline, $+$ represents the positive part of the argument and s_i takes the value ± 1 to represent the left/right orientation of the corresponding spline function. The variables to be included and their degree of smoothness are selected adaptively, as are their interactions (which can be of any order) and their degree of smoothness. As a brief explanation, suppose that the current model has I basis functions (the constant function is always included) some of which are univariate bases and some of which are multivariate tensor bases. A new basis is selected by searching among all univariate bases and selecting that which most benefits the existing model. Then the tensor products of the selected basis with each of the bases in the model are evaluated and the best tensor product is added to the model as the $(I + 1)$st basis. Once the basis selection process is complete it is followed by a similar basis deletion process in which redundant bases are removed. The remaining model bases may then be converted to piecewise cubic bases, if desired, before their use in the final model.

Friedman's MARS uses a model selection criterion based on generalized cross-validation (GCV)[32] to determine whether a candidate basis function should be added to the existing model. Criteria based upon GCV attempt to minimize the residual sum of squares adjusted for the amount of fitting being done by the model, i.e., for the increased variance associated with increased model complexity. The original GCV criterion may be stated as

$$GCV(g) = \{RSS_g/n\}/\{(1 - \mathrm{tr}(\mathbf{S})/n)^2\}$$

where n is the number of observations, RSS_g is the residual sum of squares from the fit of the model g to the data and

S is the smoothing matrix corresponding to g. However, adaptively selected knots involve more fitting and hence a higher estimation cost relative to nonadaptively selected knots. This, as well as the empirical observation that the flexibility of adaptive splines often leads to substantial overfitting[24], motivates the desire to adjust the above criterion to a statistic of the form

$$GCV(g_m) = \{RSS_{gm}/n\}/\{(1 - (m_1 + (m_2{*}d))/n)^2\}$$

as used in MARS modeling where m_2 of the m basis functions of which g is composed are adaptively selected, $m = m_1 + m_2$. The hope is that by utilizing an adjusted criterion a better balance between model variance and bias can be achieved, and that overparametrization, which is difficult to diagnose, may be avoided.

In terms of scalability Friedman provides an upper bound on the complexity of MARS of

$$nNM_{\max}^3(\alpha + \beta M_{\max}/L)$$

where α and β are constants of proportionality, n is the number of predictor variables, N is the sample size, $M_{\max}$ is the maximum number of basis functions and L is the minimal number of observations between each knot. Since MARS is implemented after dimension reduction our main concern is not the dimension of the predictor space but the sample size N. The current owners of the commercial MARS software[33] remark that MARS performs all of its training in RAM and provide limits on the number of predictor variables given the sample size and available RAM. For example, with 100 000 observations and 128 MB of RAM and the restriction that the maximum number of basis functions equal the number of specified predictors, the maximum number of predictors allowed is 45. No run time statistics are provided.

In our test bed analysis the MARS algorithm was implemented with a maximum of two variables per basis function (two-way interactions) and the remaining parameters set at their default values (df = GCV cost term = 3, fv = (fractional) incremental penalty for increasing the number of variables in the mars model = 0, ms = minimum span (minimum number of observations between each knot) = 0). Interactions between variables i and j for all i, j were permitted and MARS model response estimates for a set of covariate vectors were calculated by a call to the subroutine fmod

APPENDIX B. UTV DECOMPOSITION

Introduced by Stewart,[7,34] the UTV decomposition of a matrix **X** is of the form

$$\mathbf{X} = \mathbf{U}\begin{pmatrix}\mathrm{T}\\0\end{pmatrix}\mathbf{V}^{\mathbf{T}}$$

where the matrix **T** is upper triangular (in which case **T** is denoted as **R**) or lower triangular (in which case **T** is denoted as **L**). The UTV decomposition falls somewhere between the SVD and the QR decompositions; by using various algebraic tools and essentially sacrificing the diagonal structure of Λ in the SVD we obtain a decomposition that is faster, cheaper and easier to downdate. For example, suppose our matrix is $n \times p, n \geq p$ with rank r, e.g., our data set has 100 000 compounds and 10 000 descriptors. Then a UTV decomposition would require $\mathcal{O}(p^2)$ flops plus $\mathcal{O}((p - r)p^2)$ flops if $n \approx p$ or plus $\mathcal{O}(rp^2)$ flops if $n \gg p$.[35] In contrast, the SVD would involve $14np^2 + 8p^3$ flops if $n \approx p$ and $6np^2 + 20p^3$ flops if $n \gg p$.[6] If a unit-rank updating is necessary the number of flops required for the UTV decomposition is $\mathcal{O}(p^2)$ as opposed to $\mathcal{O}(p^3)$ flops for the SVD[7].

The UTV is an accurate approximation of the SVD provided that the off-diagonal block of **T** is sufficiently small in norm. In other words, if the effective rank of **X** is m and we let $\lambda_i(\mathbf{X})$ represent the ith largest singular value of **X**, then a ULV decomposition will yield a matrix **T** of the form

$$\begin{pmatrix}\mathrm{L_m} & 0\\ \mathrm{H} & \mathrm{E}\\ 0 & 0\end{pmatrix}$$

where $\mathbf{L_m}$ is of order m and $||(\mathbf{H},\mathbf{E})|| = \mathcal{O}(\lambda_{m+1})$ (a URV decomposition can be described similarly). This last property[6] ensures that the component matrices of **U** and $\mathbf{V^T}$ well approximate the associated subspaces of **X**, i.e., the smaller the off-diagonal elements, the better the approximation.

A ULV decomposition of the compound descriptor matrix for a selected "fitting" CV data set yields a new set of PCs as the columns of

$$\mathrm{U}\begin{pmatrix}\mathrm{L}\\0\end{pmatrix}$$

from which 12 PCs were selected as predictors in our test bed CV model fitting procedure. Computational routines for implementing the UTV algorithm were obtained from the package UTV Tools.[36]

REFERENCES AND NOTES

(1) Gershon, D. Structural genomics − from cottage industry to industrial revolution. *Nature* **1999**, *408*, 273−275.
(2) Hagmann, M. Computers aid vaccine design. *Science* **2000** *290*(6), 80−82.
(3) Abbott, A. Functional genomics: Structures by numbers. *Nature* **1999**, *408*, 130−132.
(4) West, M.; Nevins, J.; Marks, J.; Spang, R.; Zuzan, H. *Bayesian Regression Analysis in the 'Large p, Small n' Paradigm with Application in DNA Microarray Studies*; Technical Report 00-22; Duke University: Durham, NC, 2000.
(5) Dudoit, S.; Fridlyand, J.; Speed, T. *Comparison of discrimination methods for the classification of tumors using gene expression data*; Technical Report #576; Department of Statistics, University of California at Berkeley: 2000.
(6) Golub, G.; Van Loan, C. *Matrix computations*, 3rd ed.; John Hopkins University Press: Baltimore, MD, 1996.
(7) Stewart, G. An updating algorithm for subspace tracking. *IEEE Trans. Sig. Processing* **1992**, *40*, 1535−1541.
(8) Friedman, J. Multivariate adaptive regression splines. *Ann. Stat.* **1991**, *19*(1), 1−141.
(9) Wold, S. Cross validatory estimation of the number of components in principal components models. *Technometrics* **1978**, *20*, 397−406.
(10) Myers, R. *Classical and modern regression with applications*; PWS-Kent Publishing: Boston, MA, 1990.
(11) Baroni, M.; Costantino, G.; Cruciani, G.; Riganelli, D.; Valigi, R.; Clementi, S. Generating optimal linear PLS estimations (GOLPE): An advanced chemometric tool for handling 3D-QSAR problems. *Quant. Struct.-Act. Relat.* **1993**, *12*, 9−20.
(12) Rusinko, A.; Farmen, M.; Lambert, C.; Brown, P.; Young, S. Analysis of a large structure/biological activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, *9*(6), 1017−1026.
(13) Carhart, R.; Smith, D.; Venkataraghavan, R., Atom pairs as molecular features in structure−activity studies: Devinition and applications. *J. Chem. Comput. Sci.* **1985**, *25*, 64−73.
(14) Bellman, R. *Adaptive control processes: A guided tour*; Princeton University Press: Princeton, NJ, 1961.

(15) Stewart, G. *Introduction to matrix computations*; Academic Press: San Diego, CA, 1973.
(16) Vandeginste, B.; Massart, D.; Buydens, L.; De Jong, S.; Lewi, P.; Smeyers-Verbeke, J. *Handbook of Chemometrics and Qualimetrics: Part B*; Elsevier: Amsterdam, The Netherlands, 1998; Data Handling in Science and Technology − Volume 20B.
(17) Healy, M. *Matrices for statistics*; Clarendon Press: Oxford, U.K., 1986.
(18) Rencher, A. *Methods of multivariate analysis*; John Wiley & Sons: New York, NY, 1995.
(19) Fierro, R.; Hansen, P.; Hansen, P. *UTV Tools: Matlab Templates for Rank-Revealing UTV Decompositions*; Technical Report IMM-REP-99-2; Technical University of Denmark: Lyngby, Denmark, 1999.
(20) Hastie, T.; Tibshirani, R. *Generalized additive models*; Number 43, Monographs on statistics and applied probability. Chapman and Hall/CRC: Boca Raton, FL, 1990.
(21) Pittman, J. Adaptive splines and genetic algorithms. *J. Comput. Graph. Stat.*, accepted for publication.
(22) Hastie, T.; Tibshirani, R.; Friedman, J. *The elements of statistical learning: Data mining, inference, and prediction*; Springer-Verlag: New York, 2001.
(23) StatLib. http://lib.stat.cmu.edu (accessed November 1997).
(24) Ye, J. On measuring and correcting the effects of data mining and model selection. *J. Am. Stat. Assoc.* **1998**, *93*(441), 120−131.
(25) Simonoff, J. *Smoothing methods in statistics*; Springer-Verlag: New York, 1996.
(26) van der Voet, H. Comparing the predictive accuracy of models using a simple randomization test. *Chem. Intel. Lab. Sys.* **1994**, *25*, 313−323.
(27) Wehrens, R.; van der Linden, W. Bootstrapping principal regression models. *J. Chem.* **1997**, *11*, 157−172.
(28) Ihaka, R.; Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Computat. Graphical Statistics* **1996**, *5*(3), 299−314.
(29) Kaufman, L.; Rousseeuw, P. *Finding groups in data: An introduction to cluster analysis*; John Wiley & Sons: New York, 1990.
(30) Ross, D.; Scherf, U.; Eisen, M.; Perou, C.; Rees, C.; Spellman, P.; Iyer, V.; Jeffrey, S.; Van de Rijn, M.; Waltham, M.; Pergamenschikov, A.; Lee, J.; Lashkari, D.; Shalon, D.; Myers, T.; Weinstein, J.; Botstein, D.; Brown, P. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* **2000**, *24*, 227−235.
(31) Easterling, R. Quantifying the uncertainty of computational predictions. *Proceedings*; 2001 SEM Annual Conference, Portland, OR, June 4−6, 2001.
(32) Craven, P.; Wahba, G. Smoothing noisy data with spline functions. *Numerische Mathematik* **1979**, *31*(4), 377−403.
(33) Mars White Paper, Salford Systems, Inc., http://www.salford-systems.com/whitepaper.html#mars (accessed April 2001).
(34) Stewart, G. Updating a rank-revealing ULV decomposition. *SIAM J. Matrix Anal. Appl.* **1993**, *14*, 494−499.
(35) Hansen, P.; Yalamov, P. *Computing symmetric rank-revealing decompositions via triangular factorization*; Technical Report IMM-REP-2000-4; Technical University of Denmark: Lyngby, Denmark, 2000.
(36) Fierro, R.; Hansen, P.; Hansen, P. UTV tools, version 1.0. California State University: San Marcos, CA, and Technical University of Denmark: Lyngby, Denmark. http://eivind.imm.dtu.dk/as/ or http://www.netlib.org/numeralgo (accessed February 2001).

CI0103828