Local Lazy Regression: Making Use of the Neighborhood to Improve QSAR Predictions

Rajarshi Guha,*[,†,§] Debojyoti Dutta,[‡,§] Peter C. Jurs,[†] and Ting Chen[‡]

Department of Chemistry, Pennsylvania State University, University Park, Pennsylvania 16802, and
Department of Computational Biology, University of Southern California, Los Angeles, California 90089

Received February 27, 2006

Traditional quantitative structure−activity relationship (QSAR) models aim to capture global structure−activity trends present in a data set. In many situations, there may be groups of molecules which exhibit a specific set of features which relate to their activity or inactivity. Such a group of features can be said to represent a local structure−activity relationship. Traditional QSAR models may not recognize such local relationships. In this work, we investigate the use of local lazy regression (LLR), which obtains a prediction for a query molecule using its local neighborhood, rather than considering the whole data set. This modeling approach is especially useful for very large data sets because no a priori model need be built. We applied the technique to three biological data sets. In the first case, the root-mean-square error (RMSE) for an external prediction set was 0.94 log units versus 0.92 log units for the global model. However, LLR was able to characterize a specific group of anomalous molecules with much better accuracy (0.64 log units versus 0.70 log units for the global model). For the second data set, the LLR technique resulted in a decrease in RMSE from 0.36 log units to 0.31 log units for the external prediction set. In the third case, we obtained an RMSE of 2.01 log units versus 2.16 log units for the global model. In all cases, LLR led to a few observations being poorly predicted compared to the global model. We present an analysis of why this was observed and possible improvements to the local regression approach.

1. INTRODUCTION

Traditional quantitative structure−activity relationship (QSAR) models have been built using a wide variety of modeling techniques ranging from multiple linear regression and partial least squares to neural networks and random forests. A common feature underlying models built using these techniques is that they consider the entire training data set for building those models. That is, the structure−activity relationships are based on all of the molecules in the data set. We term such models *global* models. In many cases, this approach is perfectly acceptable as there may be one or more distinct structural features whose absence or presence is clearly related to the activity (or inactivity) of the molecules in the data set. Models based on such features are able to characterize the structure−activity trends in data set in general but can be unduly influenced by molecules at the extremes of the activity range. Also, when the relationship between the activity and the features is not overly complex, such global methods are very useful.

However, in some cases, the structure−activity relationship is so complex that a single model cannot characterize it fully. For example, consider the case where certain molecules within a data set may have extra features that influence their activity.[1] If such features are relevant to a small subset of molecules within the data set, a global model may not necessarily use these features to encode a general SAR. That is, the minority features may be overshadowed by the effects of more global features. As a result, methods which can focus on structure−activity trends that are not necessarily global can be used to build *local* QSAR models.

Intuitively, one can develop a local QSAR model by considering subsets of a larger data set that have similar compounds and then develop statistical models on the subsets. These subsets could be obtained by a clustering procedure. By definition, the molecules in a given cluster would be similar to each other (in the feature space used to perform the clustering), and as a result, a modeling algorithm would have a better chance of capturing a SAR that characterizes these molecules accurately. This approach has been shown to work well[2−4] but has a number of disadvantages including the fact that most clustering algorithms require that the number of clusters be specified a priori.

One aspect of the clustering-based approach to local QSAR modeling is that, given a data set which results in two or more distinct clusters, one would probably not attempt to develop a global QSAR model to characterize the data set. On the other hand, if a data set cannot be divided into distinct clusters, a local QSAR approach based on clustering may not lead to significant improvements. In this type of situation, a piecewise approach can be applied. This approach is generally used to develop linear models for data sets in which the global SAR is nonlinear in nature. Thus, by focusing on a subset of the data, one may expect that individual subsets will exhibit a linear SAR. An example of such an approach was described by Barakat et al.[5] in which they used a used a genetic algorithm to obtain subsets that exhibited linear trends. Shen et al.[6] developed piecewise linear models by performing clustering using a spanning tree and then applying a particle swarm optimization scheme to group different portions of the tree to obtain groups of molecules amenable to linear modeling. Another approach to piecewise linear

* Corresponding author phone: (814) 865-7402; e-mail: rxg218@psu.edu.
[†] Pennsylvania State University.
[‡] University of Southern California.
[§] These authors contributed equally to this work.

10.1021/ci060064e CCC: $33.50 © 2006 American Chemical Society
Published on Web 05/19/2006

modeling is the use of Kriging,[7] which assumes that the distribution of errors is not independent. Fang et al.[8] used this technique and showed that Kriging was able to improve the root-mean-square error (RMSE) by 18% compared to partial-least-squares regression. However, Kriging does require that suitable basis functions be chosen as well as splitting of the data set into representative training and testing subsets.

In this work, we consider a simplified, but different, form of the piecewise linear modeling technique in which, rather than performing a priori clustering of the data set, we determine the neighborhood of a query molecule in the data set on the fly and use that neighborhood to build a linear model, which is then used to predict the activity of the query molecule. This approach is termed local lazy regression (LLR) and has been described by Birattari et al.[9] and Bontempi et al.[10] The term *lazy* arises because of the fact that no model is built a priori. This approach is also an example of an instance-based learning technique.[11]

The lazy technique has been applied to the analysis of time series, where events are correlated in time. In the context of cheminformatics, this approach could be useful in areas such as high-throughput screening (HTS) where the initial training set can be very large and, rather than rebuilding models with the addition of new molecules, performing on-the-fly predictions using local neighborhoods can be advantageous. Though this modeling technique can be used to build both linear and nonlinear local models, we restricted ourselves to the use of linear models in this work because we compare the performance of the lazy method on data sets for which we had previously developed traditional multiple linear regression models.

2. METHODOLOGY

In this section, we present a very brief outline of local lazy regression. A more mathematical treatment of the various aspects of local regression can be found in refs 12 and 13.

2.1. Local Lazy Regression. We assume that the dependent variable, y, varies as a function of the predictor variables, x. Given a training set, a regression method can be used to construct a curve (or a line) that is used to then predict the value of the dependent variable, $\hat{y}$, for a new observation based on the values of the predictor variables, x', for the new observation. Linear regression is a very common statistical technique that finds the best possible line that minimizes the sum of squares of the residual error. Mathematically, the line is described by

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{\mathrm{p}} x_{\mathrm{p}}$$

where $\hat{y}$ is the predicted value of the dependent variable; x_1, ..., x_{p} are the p predictor variables; and β_0, ..., β_{p} are the estimated regression coefficients. We choose β_0, ..., β_{p} such that we minimize

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

where y_i and $\hat{y}_i$ represent the value of the dependent variable and the predicted value of the dependent variable for the ith observation, respectively. It can be shown that the well-known least-squares regression algorithm minimizes the squared residuals. The resultant estimated regression coefficients are given by

$$\beta = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$$

where $\mathbf{X}$ is a matrix of predictor variables, $\mathbf{Y}$ is a column vector representing the dependent variable, and β is a column vector of regression coefficients. Note that, in general, $\mathbf{Y}$ can be a matrix of dependent variables. However, in this study, we focus only on a single dependent variable.

If the relationship between y and x is very nonlinear in nature, linear regression will not result in an accurate predictor, when considering the data set as a whole. In other words, the ideal regression line will be inaccurate because of the nonlinear nature of the inherent relationship. However, it is usually possible to approximate most complicated nonlinear curves by a piece-wise approximation, namely, a collection of small lines.

We assume that we can approximate the nonlinear relationship between the dependent variable and the predictor variables by a locally linear relationship. We consider a training set observation whose predictor variables are denoted by x. Then, we assume that the relationship between the dependent variable and the predictor variables for this observation, for a small region around x, is linear. Thus, if we could determine the points around x [say NN(x)] and build a regression model with only the points in NN(x) using the least-squares method, then such a model would minimize the squared residuals for that region. Thus, we call this method a local method. However, building locally linear models would be very cumbersome for all points in the training data. Furthermore, the use of such a set of models would entail that we search through all of the models to find the one that was most suitable for a given query point. To avoid these problems, we only build a linear regression model using the local neighborhood when asked to predict a value for a query point. Thus, we do not need to build multiple models for each point in the training set beforehand. In essence, when faced with a query point, the approach builds a predictive model on the fly. Hence, this approach is termed local lazy regression.

Briefly, for a given query point, x', we first calculate the k nearest neighbors (k-NN) to form the neighbor set NN(x'). Then, we build a linear regression model with only the points in NN(x') and their corresponding y. That is, we find

$$\beta_{x'} = (\mathbf{X}^{\mathrm{T}}_{\mathrm{NN}(x')}\mathbf{X}_{\mathrm{NN}(x')})^{-1}\mathbf{X}^{\mathrm{T}}_{\mathrm{NN}(x')}\mathbf{Y}_{\mathrm{NN}(x')}$$

where $\mathbf{X}_{\mathrm{NN}(x')}$ and $\mathbf{Y}_{\mathrm{NN}(x')}$ are the values of the independent and dependent variables, respectively, for the molecules in the neighborhood of the query point. This method has several advantages over global linear regression:

- We can model complex relationships by approximating them with a collection of less complex (linear in our case) relationships.
- By deferring the model building, we avoid extra training time. This is particularly useful for large training sets.
- We store the training data explicitly, and hence, the information is never discarded.

- Because we select a fixed k for finding near neighbors, our training set is small compared to the entire global training data. Thus, the least squares algorithm runs very fast. It can be shown that this step can be done in $O(1) + O$(time for a k-NN query) time. The time to do a near-neighbor query can be reduced by using spatial data structures such as kd trees[14] or MVP trees[15] or by using approximate nearest-neighbor algorithms.[16]

Like every method, the lazy approach has a number of shortcomings. First, as all of the computations are done at query time, the determination of the local neighborhood must be efficient. Second, uncorrelated features might result in errors in the identification of near neighbors. Finally, it is nontrivial to integrate feature selection in this framework. These issues are discussed in more detail in section 5.

2.2. Determining the Neighborhood and Building a Model. It is clear that any improvement in prediction from local lazy regression is dependent on the nature of the neighborhood obtained for a given query point. The fundamental assumption here is that molecules that are close to each other in a descriptor space will exhibit similar properties.[17] The simplest approach to determining the neighborhood is to consider the k nearest neighbors in the descriptor space being considered. In general, k must be specified beforehand. We used the lazy package available for R 2.2.0 [18] to perform modeling, in which k is automatically determined using a leave-one-out (LOO) cross-validation procedure. That is, a certain value of k was chosen and a model was developed using the resultant neighborhood. The model was validated using LOO cross-validation to generate a q^2 value. k was then increased and the process repeated until q^2 decreased or did not show any significant increase.

Because the aim of local regression is to consider the close neighborhood of a query point, a linear model must be built with relatively few observations. In general, for the data sets considered here, the neighborhood contained 20−30 molecules. Given the small size of the neighborhood, it is possible that linear regression will fail, because of a singular design matrix. This is a concern especially when a larger number of descriptors are used. Thus, the lazy package uses ridge regression[19] to build the linear model. In addition, to avoid an overdetermined system, we restricted ourselves to relatively few descriptors (less than 10).

3. DATA SETS

We considered three data sets for this study. The first consisted of 179 artemisinin analogues studied by Guha and Jurs.[20] The molecules were obtained from a larger data set[21] and were studied for their activity as antimalarial drugs. The previous work[20] had developed a series of multiple linear regression and neural network models. The 179 molecules were optimized for geometry using MOPAC 7.04 and the PM3 Hamiltonian. After optimization, the ADAPT software package was used to generate a a total of 301 descriptors. The descriptor pool was then reduced using identical testing and correlation testing to remove redundant and information-poor descriptors, resulting in a reduced pool of 65 descriptors.

The second data set consisted of 79 platelet-derived growth factor (PDGFR) inhibitors studied by Guha and Jurs.[22] The data set included IC_{50} values, which we converted to a negative log scale prior to modeling. The original work had developed both linear and nonlinear models, and we were interested in observing how the LLR technique performed on a relatively small data set. As above, we used MOPAC 7.04 with the PM3 Hamiltonian to optimize the geometries of these molecules. We then generated a descriptor pool and reduced it a final pool of 41 descriptors.

Table 1. Summary of the Root-Mean-Square Errors Obtained from the Global and Linear Regression Models for the Three Data Sets Considered[a]

data set		global model	local model
artemisinin	whole data set	0.86	0.62
	prediction set	0.92	0.94
PDGFR	whole data set	0.50	0.43
	prediction set	0.36	0.31
DHFR	whole data set	2.14	1.56
	prediction set	2.16	2.01

[a] The range of the dependent variable for the artemisinin data set was 2.53 log units; for the PDGFR data set, it was 2.95 log units, and for the DHFR data set, it was 17.09 log units.

The third data set consisted of 756 inhibitors of dihydrofolate reductase (DHFR) studied by Sutherland et al.[23] The original data set reported in vitro log IC_{50} values for three cases: *Pneumocystis carnii*, *Toxoplasma gondi*, and rat liver. We considered activity values for the first case. In addition, the authors noted that the activity values for this case ranged from 0.034 nM to greater than 1000 μM and that a number of observations had indeterminate activity. As a result, we excluded molecules that had activities equal to 0 as well those that had activities greater than 1000 μM. This resulted in a data set of 672 molecules. Before modeling we also transformed the reported activities to a log scale. We then evaluated a pool of 146 descriptors using MOE.[24] The descriptors included constitutional as well as topological descriptors. In addition, a number of pseudo-3D descriptors were also evaluated. This pool was then processed to remove redundant and information-poor descriptors, resulting in a reduced pool of 36 descriptors. All calculations were performed using R 2.2.0.[18]

4. RESULTS

We compared the performance of local lazy regression models and traditional global regression models on the three data sets described in section 3. For each data set, we considered the whole data set as the training set as well as the case where the data set was split into separate training and prediction sets. A summary of the RMSEs for the two cases for each data set and model type is presented in Table 1. We present a more detailed analysis of the individual data sets below.

4.1. Artemisinin Data Set. As noted previously, local lazy regression does not build a model a priori but instead develops models on the fly for each query point. Because our goal is to compare the results of predictions obtained from a single global model to those obtained from multiple local models, we rebuilt the original linear model using the whole data set. More specifically, we searched the 65-descriptor reduced pool using a genetic algorithm to obtain an optimal subset of descriptors. The best model we obtained was a four-descriptor model in which the descriptors were the same as those reported previously.[20] The model is described in Table 2. In summary, the model exhibited an

Table 2. Summary Statistics for the Four-Descriptor Linear Model Generated for the Artemisinin Data Set[a]

	estimate	std. error	*t* value	Pr(>|*t*|)
(intercept)	−60.1703	5.0321	−11.96	2×10^{-16}
N7CH	−0.2169	0.0125	−17.30	2×10^{-16}
NSB	0.2262	0.0220	10.30	2×10^{-16}
WTPT-2	27.7323	2.4810	11.18	2×10^{-16}
MDE-14	0.1154	0.0234	4.93	2×10^{-16}

[a] N7CH, number of seventh-order chains;[27] NSB-12, number of single bonds; WTPT-2, the molecular ID number[28] considering only carbon atoms; MDE-14, the molecular distance edge vector[29] considering only primary and quaternary atoms.

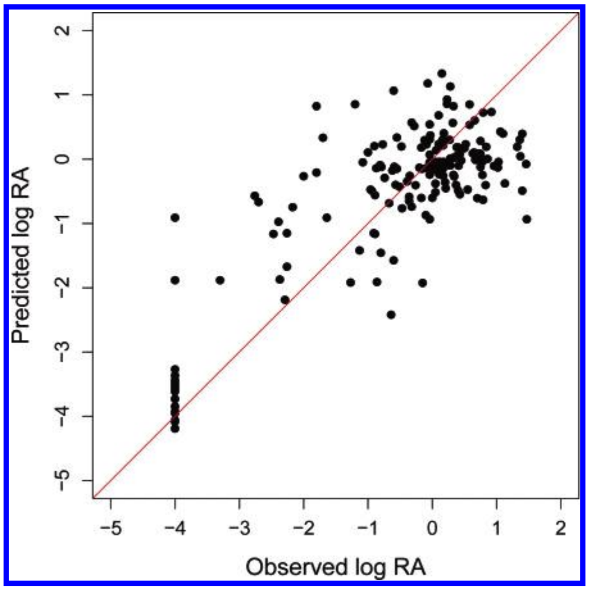


Figure 1. Plot of observed versus predicted log RA values for the artemisinin data set obtained using a four-descriptor global linear regression model. The whole data set was used to build the model.

RMSE of 0.86 (in a range of 2.53) and had an *F* value of 100.2 on four and 174° of freedom, which was greater than the critical value of 2.42 (α = 0.05). Furthermore, all of the variable inflation factors were above 1.0. Thus, the model is statistically valid. A plot of the observed versus predicted log RA values is shown in Figure 1. There are two important aspects of Figure 1, both of which indicate the poor performance of the linear model. First, there is a significant variation in the predictions for the active compounds (upper-right quadrant). Indeed, a number of the points would be considered significant outliers. Second is the group of 23 points whose observed value is −4.0 log units. Our original model[20] exhibited similar behavior. However, it is interesting to note that, even though these molecules are not predicted well, the spread of predicted values is relatively small.

Given that Figure 1 appears to indicate that there are two *groups* of molecules, one would expect that a local QSAR approach should lead to better performance. We tested the local lazy regression technique on this data set, using the descriptors obtained for the best global multiple linear regression model. This resulted in an RMSE of 0.62. The plot of predicted versus observed activity is shown in Figure 2. It is clear that, by considering local neighborhoods for each point, the regression performance increases significantly. It is also interesting to note that, out of the 23 molecules with an observed value of −4.0 log units, all but five are predicted correctly. However, compared to Figure 1, the variance in the mispredicted cases has increased.

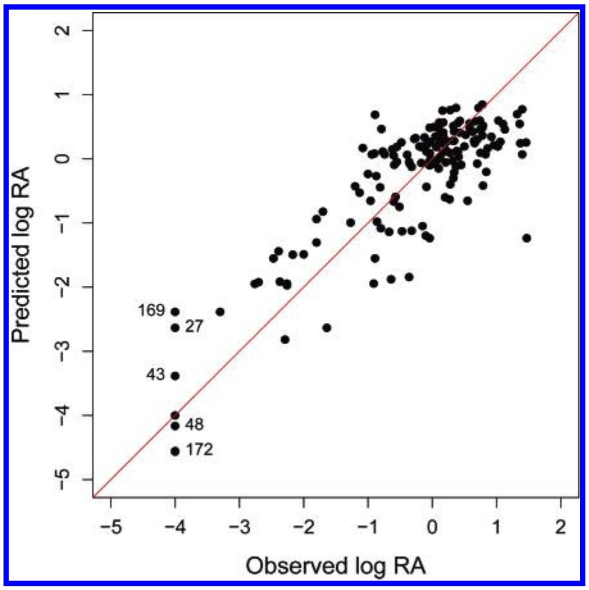


Figure 2. Plot of observed versus predicted log RA values for the artemisinin data set obtained using the four-descriptor local lazy regression model. The descriptors for this model were the same as those used for the original linear model. The local lazy regression model was restricted to linear relationships for each query point.

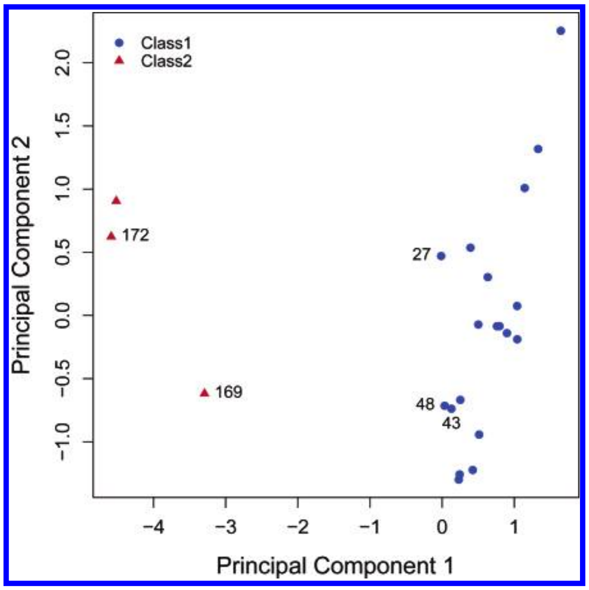


Figure 3. Plot of the first two principal components of the 23 observations with an observed activity of −4.0 log units. The cluster assignments were obtained using fuzzy analysis clustering.

Because the local lazy regression approach led to such a significant improvement in the prediction of these molecules, we considered why this might be the case. We performed a fuzzy analysis clustering[25] of the 23 molecules, using the descriptors that were used to build the local lazy regression model. A plot of the first two principal components for the 23-member subset of the original data set is shown in Figure 3. The points are colored on the basis of the cluster assignment obtained from the clustering algorithm. The number of clusters was specified a priori. However, we investigated clusterings with three and four clusters. In all such cases, one or more of the resultant clusters had silhouette values[25] which indicated the absence of any cluster structure. Thus, we restricted ourselves to two clusters. In Figure 2, we see that molecules **169** and **27** have the largest prediction error. Figure 3 provides some indication of why

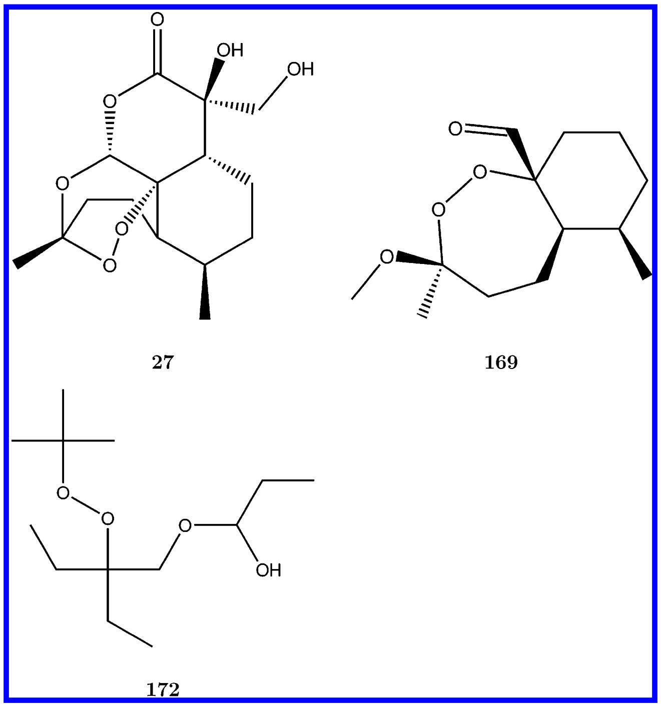


Figure 4. Structures for three molecules with an observed activity of −4.0 log units that were consistently mispredicted by the local lazy regression models.

this might be the case. It is clear that, though molecule **169** belongs to cluster 2, it is quite far from both clusters. Thus, its neighborhood (consisting of 23 points) is not really representative of the molecule. At the same time, molecule **27** is on the edge of cluster 1, and this would also explain why this molecule is mispredicted. This observation is further strengthened by noting that the silhouette width (a measure of it membership to its cluster) is less than 0.3 whereas the average silhouette width for cluster 1 is 0.83. A similar reason would explain the position of **172** in Figure 2. However, given that in Figure 3 molecules **43** and **48** are so close together, it is surprising that they have significantly different prediction errors in Figure 2. The positions of the annotated molecules in Figure 3 can be further justified by considering the neighborhoods that were used to generate the predicted values. We characterized the neighborhood by considering the mean Tanimoto similarity between all of the molecules in the neighborhood. Given that the mean Tanimoto similarity for the data set as a whole was 0.69, one would expect that, in general, neighborhoods of poorly predicted molecules would have a lower average similarity. For molecules **172**, **169**, and **27**, the average similarity of their neighborhoods was 0.46, 0.68, and 0.69, respectively. The structures of these molecules are shown in Figure 4. The fact that **169** and **172** are mispredicted is not surprising. In fact, the data set only contained two molecules that did not contain rings. This would explain the low average similarity for the neighborhood of molecule **172**. Furthermore, though **169** is polycyclic in nature, there were only a few compounds that had seven-membered rings containing a peroxide linkage. On the other hand, the ring structure of molecule **27** is quite representative of the data set, but it is the only compound that contained hydroxyl groups. This observation would also explain why the prediction errors for the molecule noted above are larger when using the local regression method compared to the errors obtained from the traditional regression method. However, the relation between prediction error and average neighborhood similarity does not always hold because the average similarity for the neighborhoods of molecules **43** and **48** are 0.80 and 0.84, respectively.

The above discussion indicates that a local regression method can improve the predictive performance of regression. However, one drawback of the above analysis is that it considered the whole data set. In general, however, one would develop a model using a training set and then use the model to obtain predictions for a set of new molecules. Thus, we randomly split the 179-molecule data set into a training and prediction set containing 129 and 50 molecules, respectively. We then rebuilt the global linear regression model on the training set, using the same four descriptors as noted above. The resultant model was used to predict the activity of the prediction set, which exhibited an RMSE of 0.92 log units. Using the local lazy regression method to obtain activity values for the prediction set led to an RMSE of 0.94. Note that we do not compare the training statistics, because the local lazy regression method does not build a model a priori. The predicted versus observed activities for the prediction set obtained using the two methods are shown in Figure 5. On the basis of RMSE values, we could conclude that the local regression approach performs poorly. However, Figure 5A shows that, though the actives are placed in a relatively small region, they do not follow the diagonal significantly. On the other hand, though the actives exhibit a higher variation in the predicted value obtained using local regression, the horizontal grouping is not very pronounced. It would thus appear that much of the increase in RMSE for the local regression is driven by the larger errors for the inactives. In fact, by ignoring molecule **127**, which has a predicted value of −5.56 log units, the RMSE for the local regression predictions drops to 0.91. If we consider the neighborhood for molecule **127**, we see that the mean distance from this molecule to the members of the training set that lay within its neighborhood was 13.59. However, the mean pairwise distance in the whole training set was 8.01. In general, poorly predicted compounds from the prediction set exhibited neighborhoods with a large average pairwise distance. However, this cannot be generalized, as there were a number of molecules whose neighborhoods were widely spread out, but which had low prediction error. In addition, corresponding behavior was observed when the average Tanimoto similarity was evaluated for the molecules in the neighborhood and compared to the average similarity for the training set as a whole. That is, neighborhoods that exhibited a low average pairwise similarity did not necessarily lead to poor predictions.

4.2. PDGFR Data Set. The original work[22] had split the data set into a training and a prediction set and then developed a linear model based on the molecules in the training set. The optimal subset of descriptors was obtained by coupling a genetic algorithm to a linear regression routine. The descriptor subset was optimized such that the resultant linear model had a low RMSE. As described above, the first step in the investigation of the LLR technique was to develop a linear model using the whole data set of 79 molecules. Because the data set is quite small, a large number of descriptors might lead to singular design matrices when building the local models. In addition, the previous work had used these descriptors to interpret the structure−activity

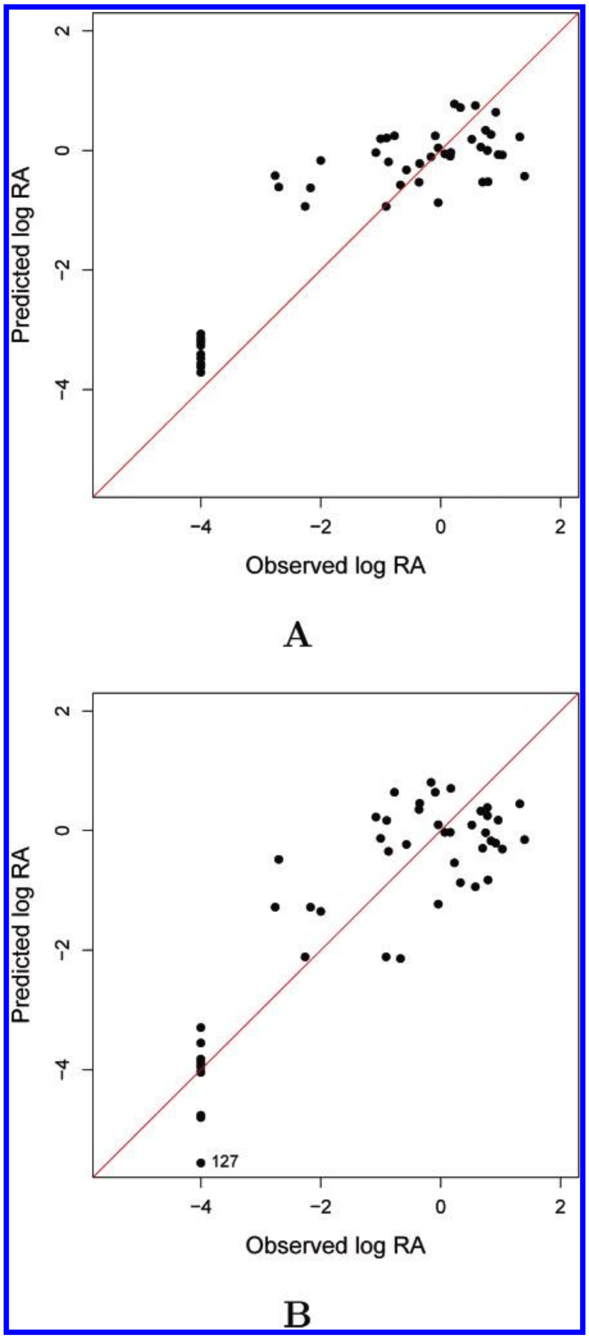


Figure 5. Predicted versus observed log RA values obtained for an external prediction set. Plot A is obtained from a global linear regression model. Plot B is obtained from a local lazy regression model.

trend. Thus, by using the same descriptors as reported previously, we hoped to identify specific deviations from previously studied structure−activity trends. The resultant model is summarized in Table 3. The model exhibited an RMSE of 0.50 log units within a range of 2.95 log units and had an *F* value of 19.76 on 3 and 75° of freedom, which was greater than the critical value of 2.73 (α = 0.05). All of the variable inflation factors were above 1.0, indicating the absence of multicollinearities. Figure 6 shows a plot of the observed versus predicted log IC_{50}'s. It is evident that, though a number of molecules in the upper right quadrant are predicted relatively well, a number of molecules toward the upper left quadrant appear to be overpredicted. Overall, it appears that the more active compounds appear to be distributed horizontally to some extent. In addition, there are two notable outliers, namely, molecules **64** and **31**.

Table 3. Summary Statistics for the Three-Descriptor Linear Model Generated for the PDGFR Data Set[a]

	estimate	std. error	*t* value	Pr(>|*t*|)
(intercept)	−1.4826	0.3601	−4.12	0.0001
MDEN-23	0.1044	0.0476	2.19	0.0315
RNHS-3	0.0399	0.0114	3.52	0.0007
SURR-5	−0.5580	0.0927	−6.02	0.0000

[a] MDEN-23, molecular distance edge vector between secondary and tertiary nitrogens;[29] RNHS-3, relative hydrophilic surface area[30] defined as the product of the sum of the hydrophilic constants and surface area of the most hydrophilic atom divided by overall log *P*; SURR-5, the ratio of atomic constant weighted hydrophobic (low) surface area to the atomic constant weighted hydrophilic surface area.[31,30]

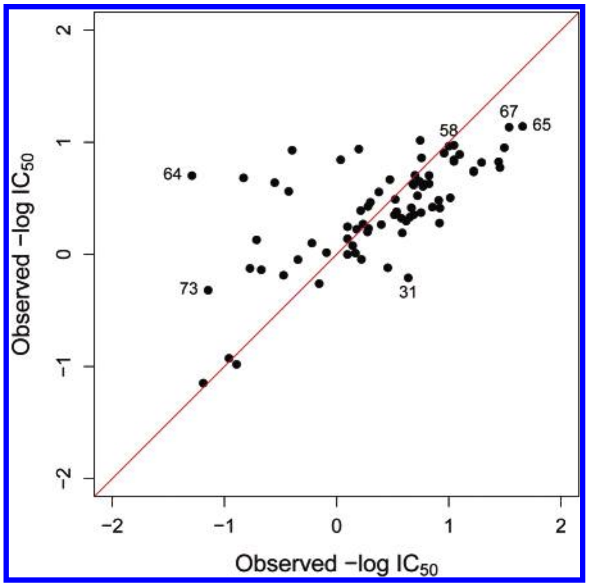


Figure 6. Plot of observed versus predicted −log IC_{50} values for the PDGFR data set obtained using a three-descriptor global linear regression model. The whole data set was used to build the model.

We then developed a local lazy regression model for the whole data set. As described above, we used the descriptors from the global linear regression model to develop the local model. The RMSE for the lazy model was 0.43, which was 12% lower than the original global regression model. Figure 7 shows a plot of observed versus predicted −log IC_{50} values obtained from the local regression model. It is evident that the local regression model leads to improved predictions as the horizontal grouping is not as evident. In addition, it leads to significantly improved predictions for certain molecules such as **73**, which was severely overpredicted by the global regression model, and **67** and **65**, which were underestimated by the global model.

Molecule **64** presents an interesting case where the local method was not able to improve the prediction by taking into account the neighborhood. Figure 8 shows the structure of molecule **64**. The LLR method determined that a neighbor containing 19 molecules was optimal. We then considered the Tanimoto similarity between molecule **64** and the members of the neighbor set. Molecule **58** (shown in Figure 8) had a similarity value of 1.0. From the structures of these two molecules, this is not surprising because they only differ in the nature of the terminal ring moiety. However, there is a significant difference in their observed activities. Further-

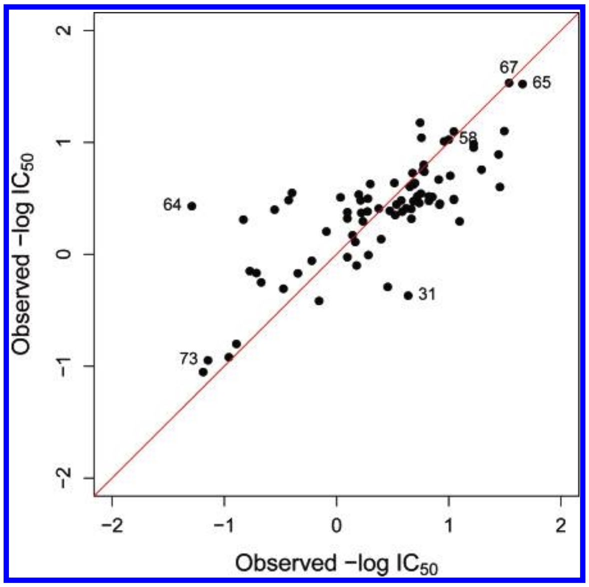
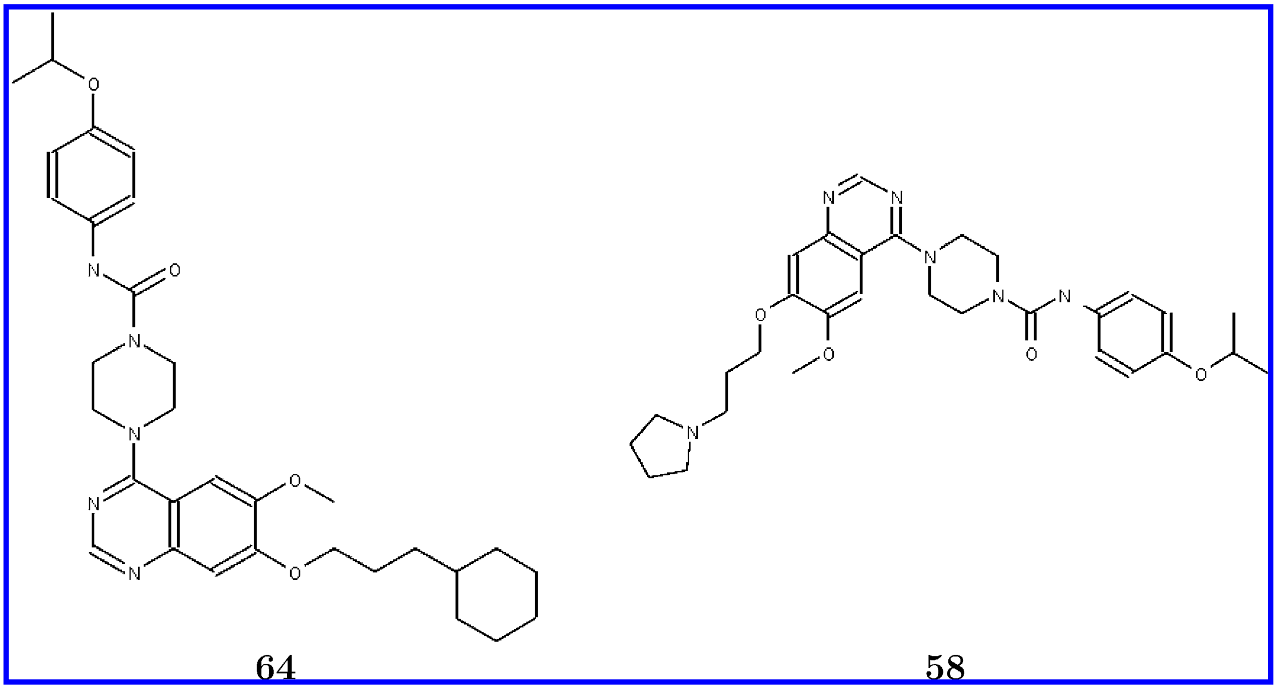


Figure 7. Plot of observed versus predicted −log IC$_{50}$ values for the PDGFR data set obtained using the three-descriptor local lazy regression model. The descriptors for this model were the same as those used for the original linear model. The local lazy regression model was restricted to linear relationships for each query point.

more, the average Tanimoto similarity of molecule **64** to the whole neighborhood is 0.94. This would indicate that the neighborhood is quite homogeneous, and a visual inspection of the neighboring molecules indicated that they shared a large common substructure. However, the observed activities of the whole neighborhood lay between 0.03 and 1.44 log units, whereas the observed activity of **64** was −1.28 log units. It is clear that, given the significant similarity in terms of structure but dissimilarity in terms of activity, a local method cannot improve the prediction of molecule **64**.

We also performed the analysis using an external prediction set. The 79-molecule data set was randomly split into a 60-member training set and a 19-member prediction set. We then built global and local regression models using the training set and then obtained predictions using the external prediction set. As described above, we do not compare the training performance of the respective models because the lazy model does not build any model a priori. The prediction set RMSE obtained from the global model was 0.36 log units. The local model exhibited a prediction set RMSE of 0.31 log units—a 14% improvement. Figure 9 displays the plots of observed versus predicted activities for the prediction sets obtained from the two models. From the plots, it is clear that the reduction in RMSE due to the local method is mainly related to the improvement in the prediction of some of the significant outliers. For example, molecules **65** and **16** are much better predicted by the local regression model. However, a number of molecules which were well-predicted by the global model (such as **34**) exhibit larger errors when predicted by the local model. At the same time, there appears to be little improvement in the prediction of molecule **75**. On inspection of the structures and neighborhood similarities of **75**, we see a similar situation as described above. The average Tanimoto similarity between **75** and the molecules in its neighborhood was 0.82. Furthermore, compared to the mean pairwise distance in the training set of 3.81, the mean distance from **75** to the molecules in its neighborhood was 0.83. These observations together with a visual inspection indicate that this molecule is structurally quite similar to its neighborhood. However, the observed activity for **75** was 0.91 log units, whereas the mean observed activity of the neighborhood was 0.56 log units. These observations highlight the fact that a local method cannot exhibit significant improvements in predictive ability when faced with structurally similar molecules which differ in their activities to a large extent. That is, a local method assumes that different groups of molecules exhibit different structure−activity trends—the key word being *structure*. If there are no significant differences in structure, this assumption fails.



64

58

Figure 8. Structure of molecules **64** and **58** from the PDGFR data set. The Tanimoto similarity between these molecules is 1.0, and it is evident that the only difference is the six-membered versus five-membered ring.

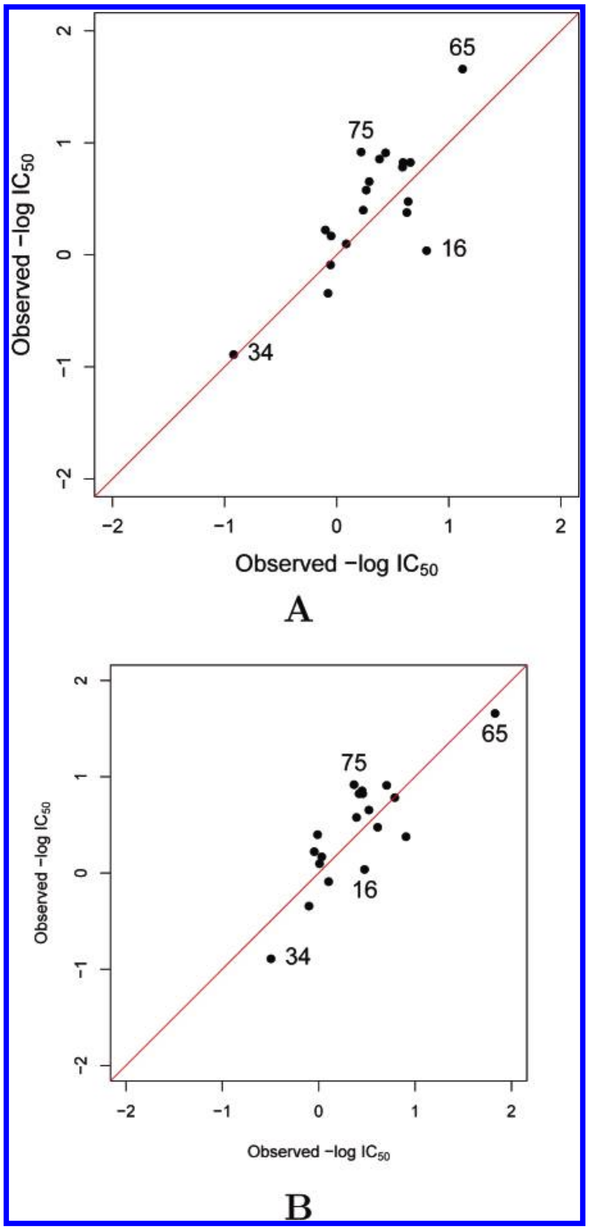


Figure 9. Predicted versus observed −log IC$_{50}$ values for the PDGFR data set obtained using an external prediction set. Plot A is obtained from a global linear regression model. Plot B is obtained from a local lazy regression model.

Table 4. Summary Statistics for the Five-Descriptor Linear Model Generated for the DHFR Data Set[a]

	estimate	std. error	*t* value	Pr(>|*t*|)
(intercept)	5.6001	0.4107	13.63	2×10^{-16}
PEOE_VSA-0	−0.0472	0.0040	−11.65	2×10^{-16}
PEOE_VSA_NEG	−0.0218	0.0031	−7.04	4.9×10^{-12}
SlogP	0.6243	0.1126	5.54	4.3×10^{-8}
SlogP_VSA1	−0.0398	0.0047	−8.46	2×10^{-16}
SMR_VSA4	0.0240	0.0055	4.36	1.5×10^{-5}

[a] PEOE_VSA-0, sum of van der Waals surface area for atoms with partial charge in the range (−0.05, 0.00); PEOE_VSA_NEG, sum of van der Waals surface area for all atoms with a negative partial charge; SlogP, log of the octanol/water partition coefficient;[32] SlogP_VSA1, sum of van der Waals surface areas for atoms whose log *P* contributions are in the range (−0.4, −0.2); SMR_VSA4, sum of van der Waals surface areas for atoms whose contributions to the molar refractivity lie in the range (0.39,0.44).

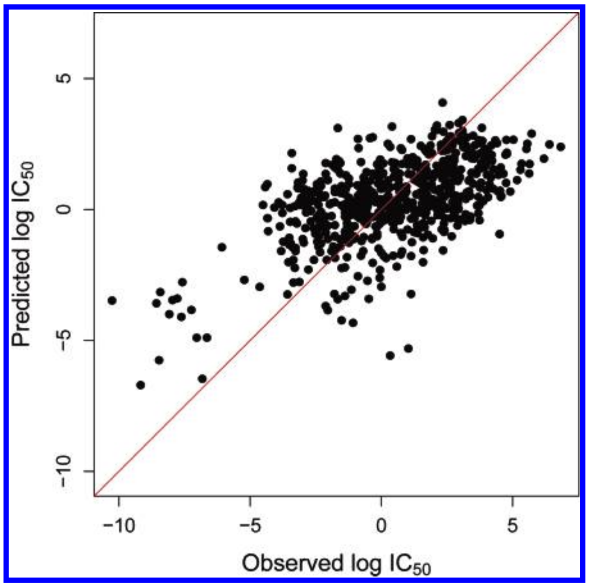


Figure 10. Plot of observed versus predicted log IC$_{50}$ values for the DHFR data set obtained using a five-descriptor global linear regression model. The whole data set was used to build the model.

4.3. DHFR Data Set. The study performed by Sutherland et al.[23] analyzed these data set using classification techniques. Because the current work focuses on regression, we developed a linear regression model using the procedure described above. We built the regression model using all 672 molecules. The best regression model was obtained by using a genetic algorithm to search the reduced descriptor pool for subsets of descriptors ranging in size from five to eight. We chose a five-descriptor model because the increase in RMSE and the overall *F* value was not significant for larger models. The statistics of the best model are summarized in Table 4. The model exhibited a RMSE of 2.14 within a range of 17.09 log units and had an *F* value of 68.7 on 5 and 666° of freedom, which was greater than the critical value of 2.27 ($\alpha = 0.05$). Furthermore, all of the variable inflation factors were above 1.0. Figure 10 shows a plot of the observed versus predicted log IC$_{50}$ values obtained by the model. The plot clearly indicates that the model does not really explain the data very well, which is further evidenced by a low R^2 value of 0.34. A number of inactive molecules are clumped together in the lower left quadrant, and the bulk of the data set (those molecules with activity greater than −5 log units) appears to have been predicted close to the mean.

We next attempted to predict the activity for the whole data set using the local lazy regression technique. As before, we used the descriptors from the initial, linear regression model as input to the local lazy regression technique. The RMSE for the local lazy regression predictions was 1.56, an improvement of 27% over the global linear regression model. The plot of the observed versus predicted activity is shown in Figure 11. It is clear that by focusing on the neighborhood of each query point the resultant predictions are significantly more accurate compared to those obtained from the global model. The predictions from the local lazy regression model exhibited an R^2 of 0.65, which was significantly better than that of the global model. The large decrease in RMSE can be explained by noting that most of the inactive molecules exhibit lower prediction errors. In addition, the bulk of the data set is now more evenly spread along the diagonal and is not clumped around the mean.

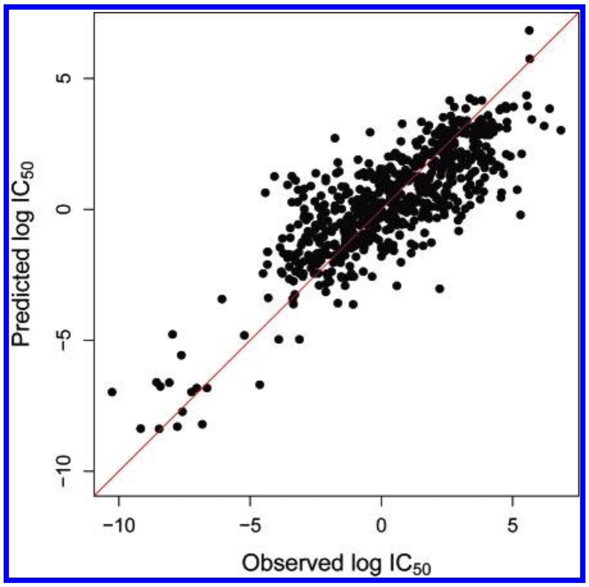


Figure 11. Plot of observed versus predicted log IC_{50} values for the DHFR data set obtained using the five-descriptor local lazy regression model. The descriptors for this model were the same as those used for the original linear model. The local lazy regression model was restricted to linear relationships for each query point.

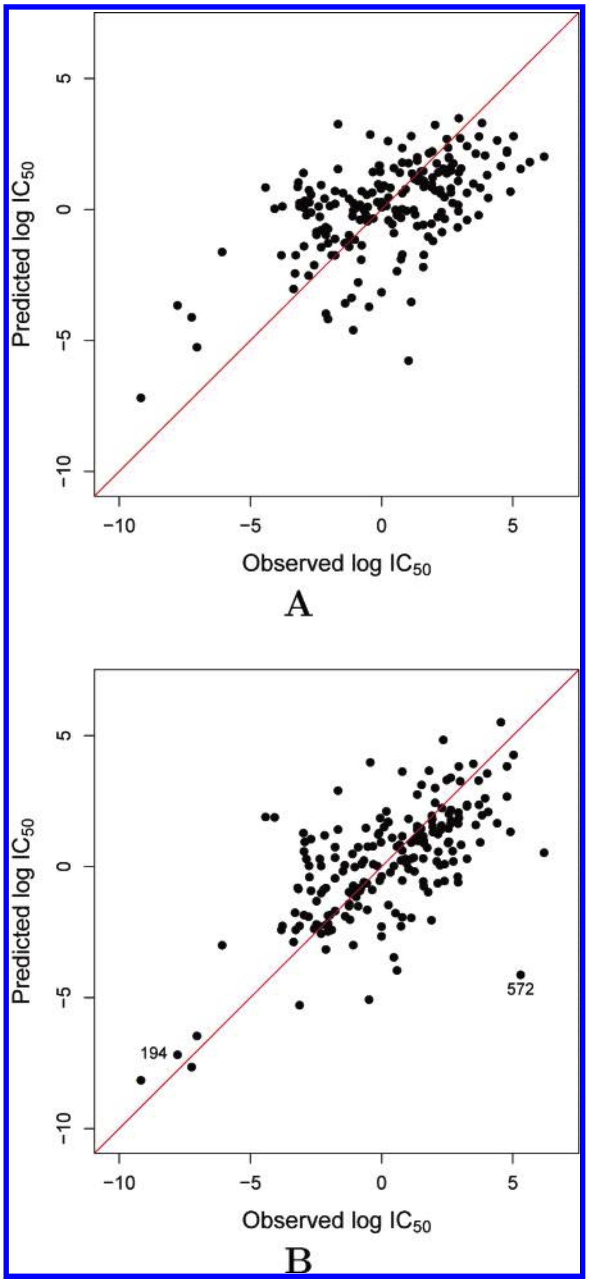


Figure 12. Predicted versus observed log IC_{50} values for the DHFR data set obtained using an external prediction set. Part A is obtained from a global linear regression model. Part B is obtained from a local lazy regression model.

As noted above, evaluating the whole data set does not really measure a technique's predictive ability. Thus, we randomly split the data set into a training set (472 molecules) and a prediction set (200 molecules). We then rebuilt the global regression model using the training set and used it to predict the activity for the molecules in the prediction set. The prediction set RMSE was 2.16 log units, and the observed versus predicted plot is shown in Figure 12A. The plot exhibits much of the same characteristics as the original global model, in that the active molecules appear to be more horizontally distributed rather than along the diagonal. Furthermore, all of the inactive molecules exhibit significant prediction errors. We then evaluated the prediction set using local lazy regression. The RMSE obtained using this technique dropped to 2.01 log units, and the plot of predicted versus observed is shown in Figure 12B. It is clear that, compared to the global regression, the results obtained from the local lazy regression are more evenly distributed. The inactive compounds now exhibit much lower prediction errors.

However, it is interesting to note that there are a few molecules which are quite severely mispredicted, such as **572**, which had a low prediction error when modeled using global regression. Intuitively, one would expect that prediction accuracy in local lazy regression would depend on the nature of the neighborhood being used to make the prediction. We compared the neighborhood of molecule **572** with that of an accurately predicted molecule (**194**) using a variety of metrics. First, we compared the average pairwise distance in the neighborhood of the two points with the average pairwise distance in the whole training set, in the space defined by the model descriptors. For the training set, this was 30.63. For molecule **572**, the average distance was 17.88, and for **194**, it was 32.18. Clearly, this is contrary to what one would expect. In fact, when the whole prediction set was considered, we observed that there was no distinct correlation between the average pairwise distance in the neighborhood for a point and its prediction error. We then considered the Tanimoto similarity of the neighborhood with respect to the overall similarity for the training set, using MACCS[26] fingerprints. The average similarity for the training set was 0.58. For the neighborhood of molecule **572**, the average similarity was 0.71, whereas for molecule **194**, this was 0.60. It is apparent that a higher average neighborhood similarity does not necessarily lead to improved prediction accuracy. As with the average pairwise distance, we observed that there was no distinct correlation between the average neighborhood similarity and the prediction set errors. However, even though the neighborhood of a molecule may exhibit a low average pairwise distance (or a high average similarity), it is important to note that the predictive accuracy is dependent on the distribution of the descriptor values in the neighborhood. When this fact is taken into consideration, the poor performance of local lazy regression on certain

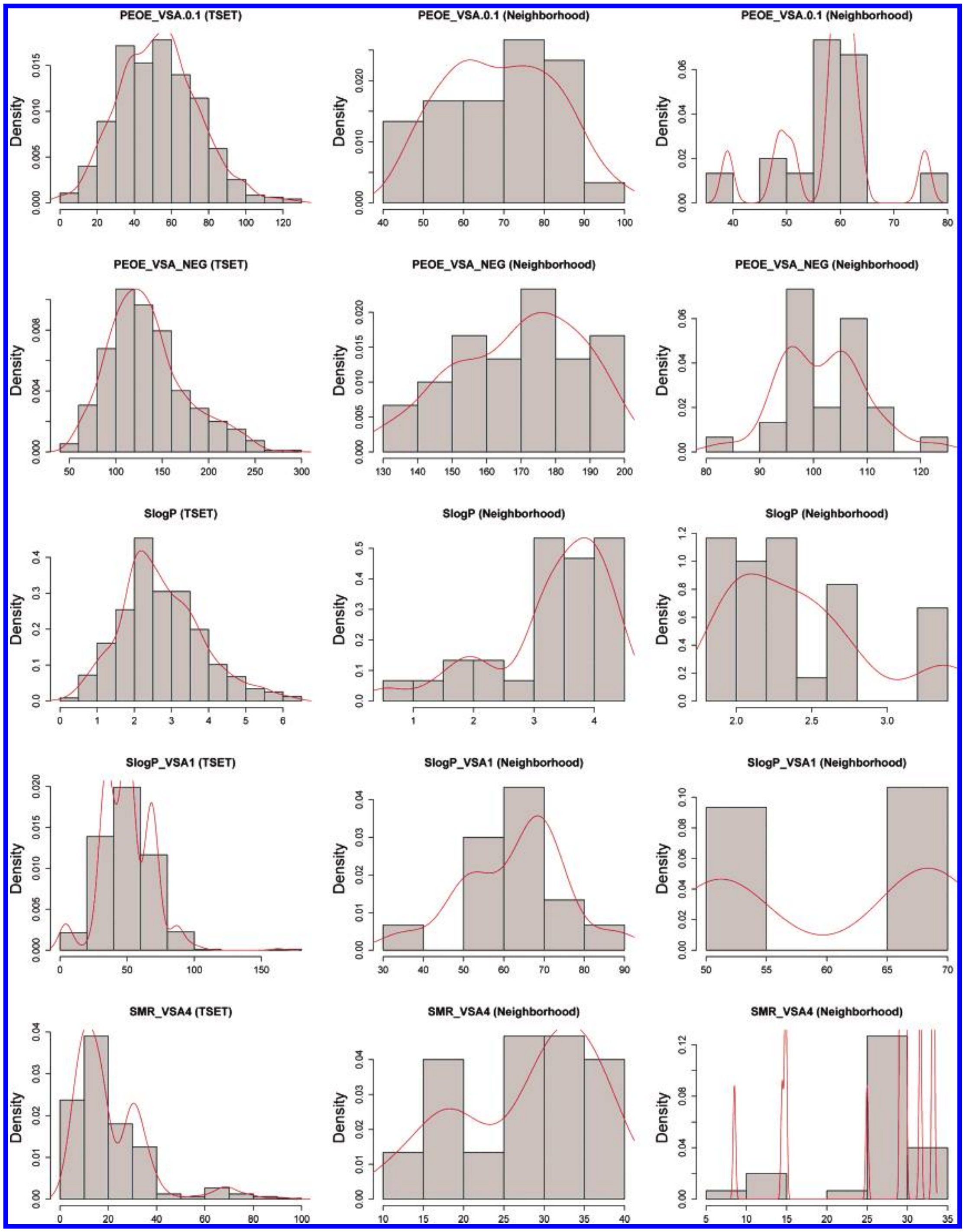


Figure 13. Histograms of the model descriptors for the training set (left) and the neighborhoods of two molecules in the prediction set. The middle column corresponds to molecule **194**, and the right column corresponds to molecule **572**. Each histogram is overlaid with the estimated density curve.

points, even though they have tight neighborhoods, can be explained. Figure 13 displays the histograms for the five descriptors used in the regression models for the training set (left column) and the neighborhoods of molecules **194** (middle column) and **572** (right column). Each histogram has the estimated density curve overlaid. If we consider the plots for the training set, we see that the density curves are close to normal distributions for the most part, though the curves for SlogP_VSA1 and SMR_VSA4 do deviate to some extent. If we then consider the histograms for molecule **194** (middle column), we see that, though the density curves do not match those for the whole training set perfectly, they are relatively representative of the training set (especially because they represent a specific subset of the training set histograms). On the other hand, if we consider the histograms for the neighborhood of molecule **572**, we see that the density curves are severely distorted, especially for the SlogP_VSA1 descriptor. Given the nature of the histograms for molecule **572**, it is not surprising that it has such a large prediction error. We investigated the histograms for other points that exhibited larger errors using local lazy regression than when using the global model, and in all cases, one or more of the density curves were significantly distorted when compared to the corresponding curves for the training set as whole.

5. DISCUSSION

The above results indicate that the use of local lazy regression can lead to an improvement in prediction accuracy, in general, but, at the same time, can also lead to larger prediction errors when compared to ordinary global regression. This is especially true when the training data is sparse. At the same time, local regression may not lead to significant improvement in predictive performance when there is no significant structural difference in groups of molecules, though they may differ significantly in terms of observed activity. One of the main advantages of local lazy regression is the fact that no a priori model need be built. This makes it suitable for large data sets, where using all of the observations can be time-consuming and even lead to overfitting. In this context, local lazy regression is suitable as a screening tool for HTS output. At the same time, because it builds a regression model for each query point, one cannot extract meaningful structure−activity trends for the data set as a whole. That is, the focus of local lazy regression is on predictive ability, rather than interpretability.

However, the use of local lazy regression opens up the possibility of using local structure−activity trends[1] rather than global trends. As we have noted previously, in most cases, a data set will have a number of features that are present in the majority of the molecules forming the basis for a global structure−activity relationship. However, it is possible that certain groups of molecules exhibit additional structural features that get *swamped* by the global trends captured by traditional models. By focusing on small neighborhoods, the possibility of exploiting local trends is significantly increased. This observation leads to one aspect of the local regression methodology that may be of concern. In the above discussion, we have built a global regression model, using descriptors selected by a genetic algorithm. Thus, these descriptors were deemed *optimal* for the global trends present in the data sets. By considering these same descriptors for use in the local regression models, we have assumed that the neighborhoods contain molecules that are similar to each other in the selected descriptor space. As a first approximation, this approach performs relatively well, though as noted above, the relation between neighborhood similarity (in terms of distances or fingerprint similarities) and prediction error is not obvious. However, it does not take full advantage of the possibility of detecting and using local structure−activity trends present in a neighborhood. That is, to really detect local structure−activity trends, one must consider some form of local feature selection, rather than relying on a fixed set of global descriptors. Thus, local feature selection would be able to detect features of the neighborhood, which may otherwise be swamped by a set of global features. But this approach leads to a fundamental problem. Without a set of initial features, how does one determine the neighborhood in the first place? One approach is to consider a set of descriptors that are not used during subsequent modeling. This set of descriptors would determine the neighborhood of a query point, which would then be analyzed using a feature selection routine to determine the best possible descriptor subset from a separate pool. Alternatively, one could use structural fingerprints. The problem with this approach is that the molecules constituting the neighborhood are close together in the initial descriptor (or fingerprint) space but are not necessarily so in the other descriptor spaces. Thus, this approach might result in no good descriptor subsets being found. An alternative approach is to use a global set of descriptors (such as that obtained from a traditional linear regression model) and use some form of stepwise selection to reduce this global set to a more focused local set of descriptors. This approach is intuitively appealing because we are able to go from a global set of descriptors to a smaller subset which is expected to focus on the structural features of the neighborhood. Both of these aspects are the focus of future investigations.

There are a number of aspects of the local lazy regression technique that need to be considered. The fundamental feature of local lazy regression is the determination of the neighborhood, given some set of descriptors. The lazy package utilizes a distance-based cross-validation approach to automatically determine the optimal number of molecules in the neighborhood of a query point. For larger data sets, this approach is not scalable as distance-based approaches have a time complexity of $O(n^2)$. The time complexity can be circumvented by the use of *k*d trees.[14] In terms of performance, the lazy package took 0.01 s to obtain predictions for the whole DHFR data set (672 molecules) running on a Pentium 4 machine with 512 MB of RAM. However, one may expect that, for data sets containing tens of thousands of molecules, traditional nearest-neighbor detection algorithms will not perform very well. In such cases, approximate nearest-neighbor approaches[16] could be utilized to rapidly determine the neighborhood. However, these approaches generally require that a radius be specified (rather than the number of nearest neighbors), thus introducing some subjectivity. At the same time, the use of a radius-based neighborhood definition may be useful, when the data set is not uniformly distributed. In such a case, the current *k*-NN approach to determining a neighborhood will always return a set of *k* molecules, even for a query compound that occupies a very sparse region of the descriptor space. In such a situation, the nearest neighbors may in fact be quite far from the query compound and, thus, may not really be representative of the query compound. In this case, performing a regression might not be valid at all. A radius-based nearest-neighbor approach would be able to point out that a query point has no (or very few) neighbors within a specified radius and, thus, rather than return an invalid prediction, would not return a prediction at all.

Another aspect of the neighborhood is that it does not always lead to a correct prediction. As has been shown above, a tight neighborhood, characterized by a low average pairwise distance or high pairwise similarity, does not correlate well with prediction error. However, it appears that the distribution of the descriptor values for the neighborhood molecules is the underlying cause of prediction error. Clearly, if we are to compare predictions from local lazy regression to those from ordinary global regression, the descriptor values for a neighborhood in local lazy regression should be representative of the whole data set. It is apparent, from Figure 13, that this is not always the case. Given this observation, one possible extension to the local regression approach is to consider a measure that takes into account the deviation of the density curves for the descriptors in a neighborhood from the density curves for the data set as a whole (or a representative sample). Predictions for query points whose neighborhood density curves exhibit significant deviations from the data set as a whole would be flagged. This approach would require some strategy to summarize multiple density curves as a single number and is currently being investigated.

Because the basis of local lazy regression is the use of the neighborhood of a query point, one alternative is to perform a clustering and determine which cluster a query point belongs to. This approach has been considered previously.[4,3] However, the disadvantage of this approach is that it is a relatively static and coarse-grained approach, because one would generally look for a number of large clusters rather than a multitude of small clusters (corresponding to the neighborhoods used in local lazy regression). Furthermore, it would require that the number of clusters be determined a priori. However, this approach is suitable for data sets which have very distinct clusters, as building a set of local models, based on cluster members, would lead to more fine-grained structure−activity relationships compared to those from a single global model.

6. CONCLUSIONS

In this paper, we have described an application of local lazy regression using linear models for individual neighborhoods. Of the three data sets tested, one did not show an improvement in the RMSE for an external prediction set, whereas the other data sets exhibited a decrease of 14% and 6% in the prediction set RMSE, when compared to traditional global linear regression models. In all cases, the use of local regression does lead to a more even distribution of points along the diagonal of a predicted versus observed plot. We have also identified why certain query points can exhibit a higher error using local regression compared to that using a global model, on the basis of the distribution of descriptor values in the query points' neighborhood, and suggest some approaches to alleviating this problem.

REFERENCES AND NOTES

(1) Sheridan, R.; Hunt, P.; Culberson, J. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* **2006**, *46*, 180−192.

(2) Pan, D.; Iyer, M.; Liu, J.; Li, Y.; Hopfinger, A. Constructing Optimum Blood Brain Barrier QSAR Models Using a Combination of 4D-Molecular Similarity Measures and Cluster Analysis. *J. Chem. Inf. Model.* **2004**, *44*, 2083−2098.

(3) Klekota, J.; Brauner, E.; Schreiber, S. Identifying Biologically Active Compound Classes Using Phenotypic Screening Data and Sampling Statistics. *J. Chem. Inf. Model.* **2005**, *45*, 1824−1836.

(4) He, L.; Jurs, P. Assessing the Reliability of a QSAR Model's Predictions. *J. Mol. Graphics Modell.* **2005**, *23*, 503−523.

(5) Barakat, N.; Jiang, J.; Liang, Y.; Yu, R. Piece-Wise Quasi-Linear Modeling in QSAR and Analytical Calibration Based on Linear Substructures Detected by Genetic Algorithm. *Chemom. Intell. Lab. Syst.* **2004**, *72*, 73−82.

(6) Shen, Q.; Jiang, J.-H.; Jiao, C.-X.; Huan, S.-Y.; Shen, G.-L.; Yu, R.-Q. Optimized Partition of Minimum Spanning Tree for Piecewise Modeling by Particle Swarm Algorithm. QSAR Studies of Antagonism of Angiotensin II Antagonists. *J. Chem. Inf. Model.* **2004**, *44*, 2027−2031.

(7) Cressie, N. *Statistics for Spatial Data*; Wiley-Interscience: New York, 1993.

(8) Fang, K.-T.; Yin, H.; Liang, Y.-Z. New Approach by Kriging Models to Problems in QSAR. *J. Chem. Inf. Model.* **2004**, *44*, 2106−2113.

(9) Birattari, M.; Bontempi, G.; Bersini, H. Lazy Learning Meets the Recursive Least Squares Algorithm. In *Advances in Neural Information Processing Systems 11*; MIT Press: Cambridge, MA, 1999; pp 375−381.

(10) Bontempi, G.; Birattari, M.; Bersini, H. Local Learning for Iterated Time-Series Prediction. In *International Conference on Machine Learning*; Publisher: Place of Publication, 1999.

(11) Aha, D.; Kibler, D.; Albert, M. Instance-Based Learning Algorithms. *Mach. Learn.* **1991**, *6*, 37−66.

(12) Cleveland, W.; Devlin, S.; Grosse, S. Regression by Local Fitting: Methods, Prospectives and Computational Algorithms. *J. Econometrics* **1988**, *37*, 87−114.

(13) Atkeson, C.; Moore, A.; Schall, S. Locally Weighted Learning. *Artif. Intell. Rev.* **1995**, *11*, 11−73.

(14) Bentley, J. Multidimensional Divide-and-Conquer. *Commun. ACM* **1980**, *23*, 214−229.

(15) Bozkaya, T.; Ozsoyoglu, M. Indexing Large Metric Spaces for Similarity Search Queries. *ACM Trans. Database Syst.* **1999**, *24*, 361−404.

(16) Dutta, D.; Guha, R.; Jurs, P.; Chen, T. Scalable Partitioning and Exploration of Chemical Spaces Using Geometric Hashing. *J. Chem. Inf. Model.* **2006**, *46*, 321−333.

(17) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(18) R Development Core Team. *R: A Language and Environment for Statistical Computing*; R Foundation for Statistical Computing: Vienna, Austria, 2005; ISBN 3-900051-07-0.

(19) Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning*; Springer: New York, 2001.

(20) Guha, R.; Jurs, P. The Development of QSAR Models To Predict and Interpret the Biological Activity of Artemisinin Analogues. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1440−1449.

(21) Avery, M. A.; Alvim-Gaston, M.; Rodrigues, C. R.; Barreiro, E. J.; Cohen, F. E.; Sabnis, Y. A.; Woolfrey, J. R. Structure Activity Relationships of the Antimalarial Agent Artemisinin. The Development of Predictive in Vitro Potency Models Using CoMFA and HQSAR Methodologies. *J. Med. Chem.* **2002**, *45*, 292−303.

(22) Guha, R.; Jurs, P. Development of Linear, Ensemble, and Nonlinear Models for the Prediction and Interpretation of the Biological Activity of a Set of PDGFR Inhibitors. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 2179−2189.

(23) Sutherland, J.; O'Brien, L.; Weaver, D. Spline-Fitting with a Genetic Algorithm: A Method for Developing Classification Structure−Activity Relationships. *J. Chem. Inf. Model.* **2003**, *43*, 1906−1915.

(24) Chemical Computing Group Inc. *Molecular Operating Environment (MOE 2004.03)*; Chemical Computing Group: Montreal, Quebec, Canada.

(25) Kaufman, L.; Rousseeuw, P. *Finding Groups in Data: An Introduction to Cluster Analysis;* Wiley: New York, 1990.

(26) MDL Information Systems Inc.

(27) Kier, L.; Hall, L. *Molecular Connectivity in Structure Activity Analysis*; John Wiley & Sons: Hertfordshire, England, 1986.

(28) Randic, M. On Molecular Idenitification Numbers. *J. Chem. Inf. Comput. Sci.* **1984**, *24*, 164−175.

(29) Liu, S.; Cao, C.; Li, Z. Approach to Estimation and Prediction For Normal Boiling Point (NBP) of Alkanes Based on a Novel Molecular Distance Edge (MDE) Vector, λ. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 387−394.

(30) Stanton, D.; Mattioni, B. E.; Knittel, J.; Jurs, P. Development and Use of Hydrophobic Surface Area (HSA) Descriptors for Computer Assisted Quantitative Structure−Activity and Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1010−1023.

(31) Mattioni, B. E. The Development of Quantitative Structure−Activity Relationship Mode ls for Physical Property and Biological Activity Prediction of Organic Compounds. Ph.D. Thesis, Pennsylvania State University, University Park, PA, 2003.

(32) Wildman, S.; Crippen, G. Prediction of Physiochemical Parameters by Atomic Contributions. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 868−873.

CI060064E