R-NN Curves: An Intuitive Approach to Outlier Detection Using a Distance Based Method

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 January 10, 2006

Libraries of chemical structures are used in a variety of cheminformatics tasks such as virtual screening and QSAR modeling and are generally characterized using molecular descriptors. When working with libraries it is useful to understand the distribution of compounds in the space defined by a set of descriptors. We present a simple approach to the analysis of the spatial distribution of the compounds in a library in general and outlier detection in particular based on counts of neighbors within a series of increasing radii. The resultant curves, termed *R*-NN curves, appear to follow a logistic model for any given descriptor space, which we justify theoretically for the 2D case. The method can be applied to data sets of arbitrary dimensions. The *R*-NN curves provide a visual method to easily detect compounds lying in a sparse region of a given descriptor space. We also present a method to numerically characterize the *R*-NN curves thus allowing identification of outliers in a single plot.

1. INTRODUCTION

A common task in cheminformatics workflows is the analysis of libraries of molecular structures. Libraries are used in a number of contexts such as virtual screening (VS) and QSAR modeling. One of the fundamental steps in the analysis of libraries is to transform or *embed* the structures into a suitable *chemical space* (also known as a descriptor space), which is a multidimensional space defined by a set of molecular descriptors. This space is subsequently mined using a variety of statistical and machine learning algorithms. As a result of the embedding process, different regions of the space can be occupied to different extents. That is, some regions may contain very few molecules, whereas other regions may be very densely populated. The analysis of a chemical library to determine the distribution of molecules in a descriptor space is termed diversity analysis.[1−3]

Diversity analysis can be leveraged in a number of scenarios. For example, during compound acquisition one would prefer to enhance the sparse regions of the library rather than add molecules to regions of the libraries' chemical space that are already densely populated. Another important application of diversity analysis is to determine which compounds lie in sparse regions of the descriptor space (termed outliers). This knowledge is useful because an active compound located in a sparse region of the chemical space might be a good lead since it is both active as well different from the bulk of the library. Thus, such outliers are good starting points for *lead hopping*.[4,5] In addition, identification of sparse regions of a library is useful since it allows one to remove outliers leading to a more uniform distribution of compounds in the given chemical space. Thus, in several scenarios, it may be important to study the distribution of compounds in a chemical space and consequently rapidly identify compounds both in sparse regions of the space (termed outliers) and, to a lesser extent, in dense regions of the space.

Determining the diversity of the chemical space has received a lot of attention in the recent past.[3,6−17] Most diversity techniques can be classed as one of two types: cell based and distance based. In the former, the descriptor space is divided into bins, and then each structure is mapped to a bin. This approach is useful for chemical spaces of low dimensions. For higher dimensional spaces, the majority of the bins will be empty resulting in a lack of discrimination, unless the number of bins is increased significantly. Thus the choice of bin size[18] can be an important factor in the success of this approach. However this approach is computationally inexpensive and well suited for low dimensional libraries.

Distance based approaches consider diversity in terms of the distances between structures in the descriptor space being considered. These approaches generally require the determination of pairwise distances. Such an approach has a time complexity of $O(n^2)$ and is not scalable for large libraries. However techniques such as the use of *k*-D trees[7] can be used to speed up these types of calculations. Another approach to circumventing the time complexity of distance based methods is the use of statistical techniques such as the KS test.[8]

Other approaches such as sphere exclusion based[14] and information theory[17] based methods have also been described. We also differentiate between methods that can be applied to arbitrary descriptor spaces and those that are based on specific descriptors[1,2] or other molecular features such as pharmacophores[15,10] or surfaces.[16]

1.1. Our Contributions. In this paper we describe a distance based method that characterizes molecules in a data

[*] Corresponding author e-mail: pcj@psu.edu.
[†] Pennsylvania State University.
[‡] University of Southern California.
[§] These authors contributed equally to this paper.

10.1021/ci060013h CCC: $33.50 © 2006 American Chemical Society
Published on Web 06/01/2006

set in a rigorous and deterministic fashion as either sparse or dense according to the nature of the corresponding regions of the descriptor space where they lie. Our method is similar to *k* NN methods,[7] which characterizes a molecule based on the number of nearest neighbors in the descriptor space. However our method is based on the number of neighbors of a query molecule located in an *N* dimensional hypersphere of radius *R* centered on the query molecule, for varying *R*. This is distinct from the traditional *k* NN approach where the *k* nearest neighbors of a query molecule are noted, however far from the query molecule they may lie. Clearly the traditional approach requires us to have a predefined value of *k*. Furthermore, in this approach the nearest neighbors might in fact be very far from the query molecules in the descriptor space. In section 5 we discuss why our approach is advantageous compared to the traditional *k* NN approach in the context of QSAR modeling. After determining the neighbor counts we then plot the neighbor count versus radius to generate an *R*-NN curve for the molecule, and we use this curve to classify the space around a compound as either sparse or dense. We show, theoretically as well as empirically, that these curves are sigmoidal in nature. These curves can then be numerically characterized to grade all the molecules in a data set in terms of their location in sparse or dense regions of the data set, allowing for easy visual identification of outliers. The naive version of the method has a time complexity of $O(n^2)$. We describe how this process can be speeded up, in sub $O(n^2)$ time, by applying an approximate nearest neighbor algorithm known as Locality Sensitive Hashing (LSH),[19] which has sublinear time complexity for each *R*-NN query, thus allowing our method to be used for large data sets.

We do not aim to provide a mechanism to select descriptors that would result in a specific distribution of molecules in the descriptor space. Rather our method can be applied to arbitrary descriptor spaces. Furthermore the method allows one to easily characterize the location of individual molecules in an intuitive manner. It also provides a global view of the distribution of molecules in the descriptor space considered. Essentially, our method can be considered a 2D approach to visualizing the distribution of molecules in a multidimensional descriptor space.

2. METHODOLOGY

2.1. Theoretical Analysis. In this section, we present a simple mathematical argument to justify our characterization of compounds based on the characteristics of the *R*-NN curve. We consider a very simple model. We consider two scenarios: (1) the query point is in a sparse region, and (2) the query point is in a dense region.

In the first scenario, we assume that around a given point, say *p*, the space is very sparse with neighbor density ρ_0. At some critical radius r_0, the neighbor density changes to ρ_1, with $\rho_1 \gg \rho_0$. Also assume ρ_0,ρ_1 to be constant. In fact it can be shown that with nonconstant densities, our result still holds true. After a radius *R*, the number of near neighbors becomes a constant, i.e., equal to the number of compounds *N*.

For a radius $r < r_0$, consider a small change in radius *dr*. Then the neighbors on the strip is $2\pi r\rho_0.dr$, which on integration, gives $2\pi\rho_0 r^2$. Similarly consider a radius $r_0 < r < R$. The number of near neighbors is now

$$2\pi\rho_0 {r_0}^2 + \int_{r_0}^{r} 2\pi r\rho_1 dr$$

Thus the number of near neighbors as a function of the radius *r*, denoted as NN(*r*), is given by

$$\mathrm{NN}(r) = \begin{Bmatrix} 0 & \text{if } r \leq 0 \\ 2\pi\rho_0 r^2 & \text{if } 0 < r \leq r_0 \\ 2\pi\rho_0 {r_0}^2 + 2\pi\rho_1(r^2 - {r_0}^2) & \text{if } r_0 < r < R \\ N & \text{is } r \geq R \end{Bmatrix}$$

The above is a piecewise quadratic function which is smooth everywhere except for the two transition radii, r_0 and *R*. From ref 20, it is clear that the above function is an approximation of the sigmoid function. This implies that for compounds in a sparse region, the *R*-NN curve can characterize the near neighbors as a function of query radius.

The argument is very similar for compounds in the dense region. The relation between the densities will be reversed. Also note that we have considered only one transition, i.e., from a sparse to a dense region or vice versa. Actually our arguments will hold for the other cases too. The characteristics will always look like a sigmoid curve due to the increasing nature of the number of near neighbors and the fixed number of total neighbors.

2.2. Generating the *R*-NN Curves. The first step in evaluating the *R*-NN curve for a given descriptor space is to determine the maximum pairwise distance in the data set. For smaller data sets this can be performed using a brute force method by evaluating the Euclidean distance matrix. For larger data sets, this is not an efficient approach and can consume large amounts of memory and time (since this calculation has a time complexity of $O(n^2)$). One approach to determining the maximum pairwise distance for a large data set is to perform random sampling multiple times. Given a sufficient number of samples, the maximum pairwise distance in these samples should approximate the exact value for the whole data set.

We then denote the maximum pairwise distance in the data set as $D_{\max}$. Next, consider the first observation in the data set. For this observation we determine the number of neighbors that lie within a sphere of radius, *R*, centered on it. We then repeat this step for different *R*. Thus, for a given observation we will generate a count of nearest neighbors within a set of specific radii. This procedure is then repeated for all the molecules in the data set if required. This procedure is summarized in pseudocode in Algorithm 1.

Clearly, an important aspect of the algorithm is the selection of radii. The choice of *R* is based on two observations. In general, at low radii a given molecule will have few neighbors, whereas for higher radii, the number of neighbors will increase. For the former case, the lower limit of *R* is 0. Thus for $R = 0$ every molecule will have zero neighbors (not including itself). For the latter case, it is clear that for $R \geq D_{\max}$ all the molecules in the data set will be neighbors for a given molecule. Thus $0 < R \leq D_{\max}$. We implemented Algorithm 1 using *R* values ranging from 1% of $D_{\max}$ to $D_{\max}$ in increments of 1%. The choice of *R* values is data set dependent. More specifically, we desire that the

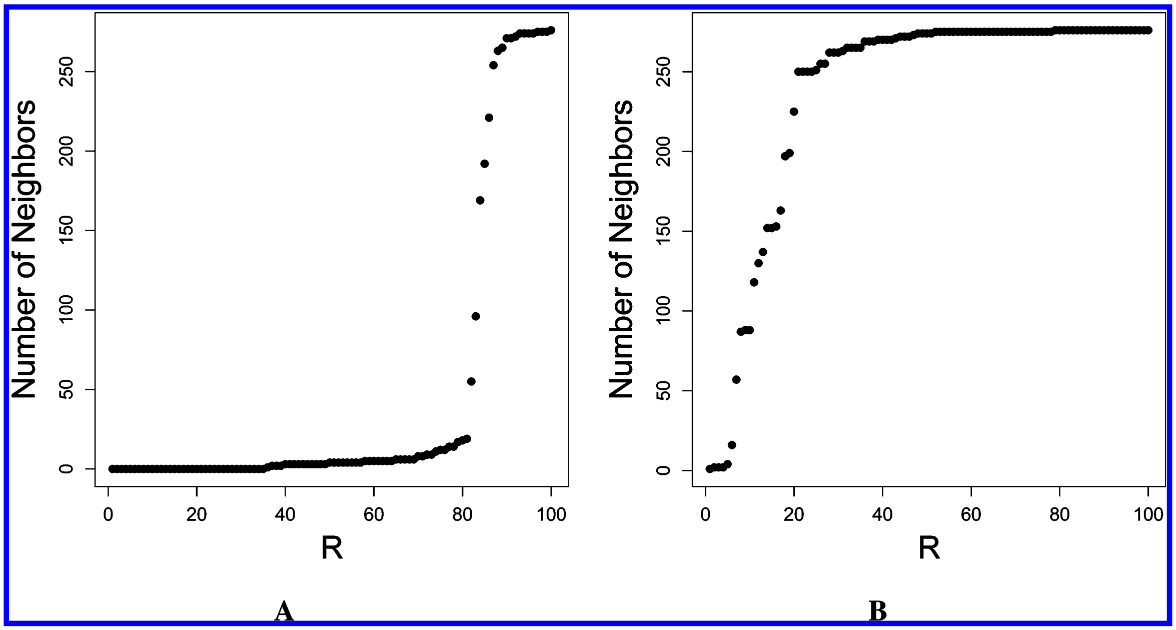
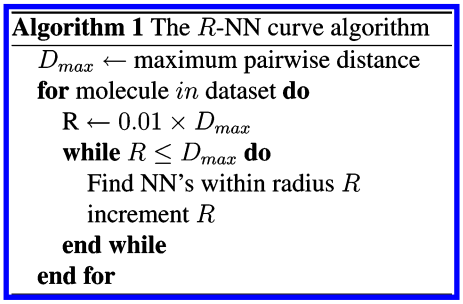


Figure 1. *R*-NN curves for two molecules lying in different regions of a descriptor space. **A** is the curve for a molecule lying in a sparse region of the descriptor space. **B** is the curve for a molecule lying in a densely populated region of the descriptor space.

R-NN curves are relatively smooth but not too detailed as to become a stepwise curve. For the data sets considered in this study the resolution of *R* values mentioned above satisfied these requirements.



Algorithm 1 The R-NN curve algorithm

```
$D_{max} \leftarrow$ maximum pairwise distance
for molecule $in$ dataset do
    R $\leftarrow 0.01 \times D_{max}$
    while $R \leq D_{max}$ do
        Find NN's within radius $R$
        increment $R$
    end while
end for
```

2.3. Characterizing the *R*-NN Curves. It was observed that the *R*-NN curves for arbitrary descriptor subsets are sigmoidal in nature. In a number of cases, the sigmoid nature of the curve is not always apparent due to very small size (or even absence) of the lower tail. However, visual inspection of the *R*-NN curves showed that every molecule in each data set exhibited the linear, exponentially increasing and saturation regions of the standard sigmoidal function. The characteristic feature of the *R*-NN curves that allows us to differentiate a molecule lying in a sparse region of the descriptor space from one lying in a dense region is the lower tail. It is observed that curves for the former class of molecules have an extended lower tail. That is, for increasing radii, the number of nearest neighbors within that radius does not increase significantly. Only after a certain *threshold radius* is reached, does the number of nearest neighbors start increasing rapidly. Visually, the difference between a molecule in a sparse region and a dense region of the data set is shown in Figure 1. Curve **A** clearly indicates that the increase in number of neighbors is very slow until $R \approx 0.5 \times D_{\max}$. That is, until the radius becomes appreciable large, the molecule has very few neighbors compared to the total number of molecules in the data set. On the other hand, **B** is the curve for a molecule that lies in a dense region of a descriptor space. In this case, the lower tail of the sigmoidal curve is nearly absent. That is, even for small radii this molecule has an appreciable number of neighbors which increases rapidly with increase in *R*.

Clearly, visual inspection of *R*-NN curves can easily allow one to distinguish between molecules in dense and sparse regions of a descriptor space. However, for larger data sets visual inspection can become unwieldy. Thus, a numerical characterization of *R*-NN curves would lead to a more efficient approach to determine whether a molecule is an outlier or not. Given our premise that *R*-NN curves are sigmoidal in nature one approach to characterization is to fit a general logistic function to the *R*-curve data for a each molecule.

The general form of the logistic function can be written as

$$N_{\mathrm{N}} = a \cdot \frac{1 + me^{-R/\tau}}{1 + ne^{-R/\tau}} \qquad (1)$$

where N_{N} is the number of neighbors within a radius *R* and *a*, *m*, *n*, and τ are the parameters of the fit. As shown above, outlying molecules will have *R*-NN curves with extended and relatively flat lower tails. Determining the value of *R* at which the lower tail transitions to the linear portion of the sigmoidal curve, we will be able to rank molecules in terms of their outlyingness. As a result we can avoid the actual nonlinear fitting procedure. Thus, we need to determine the slope of eq 1 for varying *R*. We can simplify the task by considering the value of *R* for which the slope of the sigmoidal curve is maximal. It is clear that the sigmoidal curve has a maximal slope at points along the linear portion of the curve, shown schematically in Figure 2. Since this portion of the curve is located at higher *R* values for outlying points than for points in dense regions of the descriptor space, identification of the *R* value for which the slope is maximum allows us to rank molecules as described above. Due to our choice of *R*, the task is further simplified since the slope can be calculated numerically by simply taking the difference

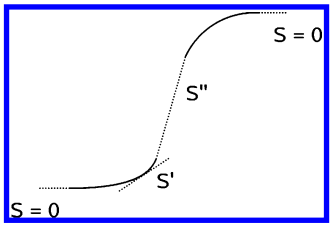


Figure 2. A schematic diagram of a *R*-NN curve highlighting the slopes, shown by dotted lines, at different points. At the initial and end points of the curve the slope is 0. The extent of the lower tail can be identified by finding the radius where the initial exponential increase begins, identified by a slope, *S*′. However since this slope varies, it is easier to identify radius corresponding to the linear portion which will have a constant slope of *S*″. Since *S*″ will have the maximum slope of the curve, it is readily identifiable.

of the neighbor counts at successive *R*. That is,

$$S_R = N_{N,R} - N_{N,R-1} \tag{2}$$

where S_R is the slope of the *R*-NN curve at a radius *R* and $N_{N,R}$ is the number of neighbors at that radius. Thus for each molecule, we evaluate S_R for $2 \leq R \leq 100$ and determine the radius, $R_{\max(S)}$, at which the maximum value of S_R occurs. Plotting the values of $R_{\max(S)}$ will allow easy visual identification of successively outlying molecules in a data set.

3. DATA SETS

We considered two data sets. The first data set consisted of 277 compounds whose measured property was boiling point.[21] This data set consisted of small molecules with a mean molecular weight of 115. The average Tanimoto similarity of the data set using MACCS fingerprints was 0.20 ($\sigma = 0.19$). MOE[22] was used to perform a geometry optimization on the structures using the MMFF94 force field. After optimization we used MOE to evaluate a set of descriptors including topological, geometric, and electronic descriptors. A total of 213 descriptors were calculated. The descriptor pool was then reduced to remove constant and correlated descriptors. Constant descriptors were identified by having a standard deviation of zero. In addition, descriptors that were constant for 80% of the observations in the data set were also removed. Next, the pairwise correlation between the remaining descriptors was evaluated. Foe each pair exhibiting a correlation (R^2) greater than 0.7, a randomly chosen member of the pair was removed from the descriptor pool. The descriptor reduction protocol resulted in a reduced pool containing 64 descriptors.

The second data set consisted of 4337 compounds studied by Kazius et al.[23] The molecules were investigated for their response to the AMES test for mutagenicity. The data set consisted mainly of small molecules though a number of larger molecules (MW > 600) were present. The average molecular weight for this data set was 240. The structures in this data set were geometry optimized in MOE using the MMFF94 force field. A total of 142 descriptors were calculated. The descriptor pool was then reduced as described above resulting in a reduced pool containing 45 descriptors. On average this data set exhibited a similar pairwise Tanimoto similarity of 0.21 ($\sigma = 0.13$) compared to the boiling point data set.

For each data set, the reduced descriptor pool was autoscaled. All calculations were performed using R 2.2.0.[24]

4. RESULTS

For the boiling point data set we had previously generated statistical models. For this data set we focused on the descriptor subset representing the best linear model, which was a 4-descriptor linear regression model. This model was obtained using a genetic algorithm to search for descriptor subsets that would generate a linear regression model with low root-mean-square error. In the case of the Kazius data set, we did not have any previous models. As a result we considered a number of randomly selected descriptor subsets. We believe that this does not detract from the analysis since our premise is that a data set will exhibit sigmoidal *R*-NN curves for any arbitrary descriptor space. Since the different random descriptor subsets resulted in similar conclusions, we only describe the results from a 5-descriptor random subset.

4.1. Boiling Point Data Set. Figure 3 displays the *R*-NN curves for 25 molecules selected randomly from the boiling point data set. The plots colored in blue represent molecules that occupy sparse regions of the descriptor space. It is evident these molecules exhibit an extended lower tail of the sigmoidal curve. As described above this indicates that for large values of *R* these molecules have few neighbors in the resultant hypersphere. In comparison most of the other curves exhibit a minimal lower tail, and the majority of those shown have no lower tail at all. Figure 4 shows a series of *R*-NN curves for molecule **226** generated using randomly selected descriptor subsets ranging in size from three to eight. In the original descriptor space this molecule was a distinct outlier, as shown in Figure 3. In one of the random descriptor spaces it is also an outlier, whereas for the other descriptor spaces shown it appears to be located in a more dense region. It is well-known that relationships in one descriptor space do not necessarily carry over to other descriptor spaces.[25] Thus, given that molecule **226** is an outlier in our original descriptor space, there is no guarantee that it is also an outlier in other descriptor spaces. However in all cases, the *R*-NN curves exhibit a sigmoidal character.

4.2. Kazius Data Set. For the Kazius data set we followed the procedures described above. A random selection of molecules from the data set exhibited *R*-NN curves similar in nature to this displayed in Figure 3. To initially identify a set of molecules lying in sparse regions of the 5-descriptor space, we considered the number of neighbors for each molecule at 50% of the maximum pairwise radius. We then considered those molecules which had less than 433 neighbors (approximately 10 of the whole data set) at this radius. In general, the choice of how many neighbors should be considered to indicate that a molecule occupies a sparse region is subjective. In the case of this data set, the mean number of neighbors at 50% of the maximum pairwise radius was 4330, that is, nearly every compound had the whole data set as its neighbors at 50% of the maximum pairwise distance. Thus molecules that would have less than 433 at this radius can be safely considered as outliers. Clearly, using values greater than 433 lead to a less stringent cutoff. At the same time, we can constrain the selection of outliers by considering a higher radius for a fixed neighbor count. In fact at 50% of the maximum pairwise radius, only two molecules **1861** and **3904** were flagged as outliers; compared to a mean number of neighbors equal to 4330 for a radius

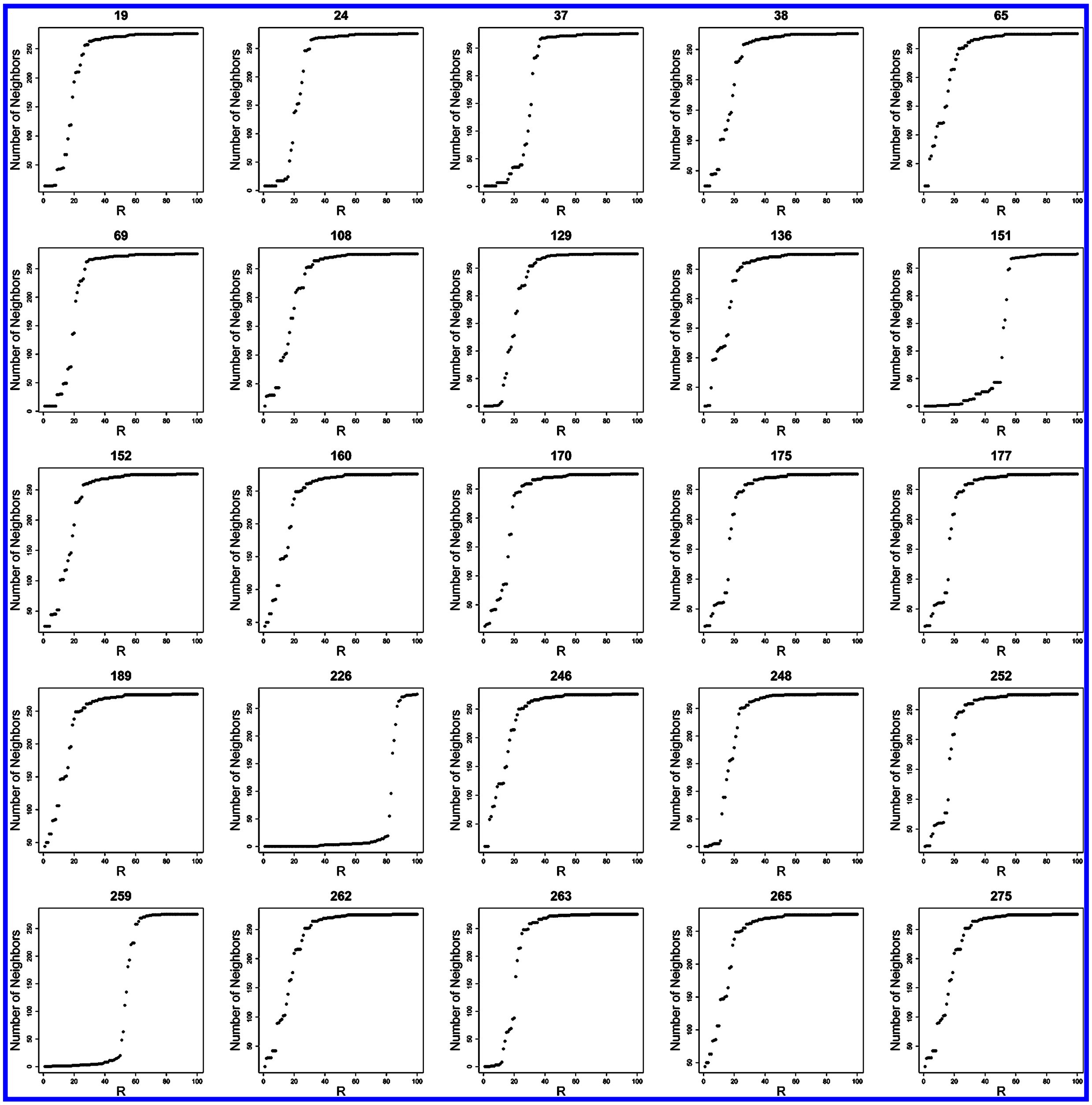


Figure 3. *R*-NN curves for 25 observations taken from the boiling point data set using four descriptors. The title of each plot indicates the serial number of the molecule in the data set. The plots in blue represent the curves for molecules lying in the sparse regions of this descriptor space. Note the extended lower tail of the sigmoid curve for these observations.

of 50% of the maximum pairwise distance, these two compounds had 41 and 2 neighbors, respectively. The structures of these compounds are shown in Figure 5, and Figure 6 shows the *R*-NN curves for these two molecules. For comparison, the structure and curve for molecule **1801** is also shown. This molecule had slightly less than half the data set (1860) as neighbors when the radius was set to 10% of the maximum pairwise distance, clearly indicating that it lies in a dense region of the descriptor space. It is clear from Figure 6 that molecule **3904** is a significant outlier. In fact at 75% of the maximum pairwise distance it only had 31 neighbors. It is also interesting to note the large value of the average number of neighbors at 50% maximum pairwise distance. Given that this value is close to the size of the data set, we conjecture that in the given descriptor space, the molecules are relatively uniformly distributed. We also considered *R*-NN curves for random descriptor subsets, using molecule **3904**. As expected, all curves exhibited sigmoidal characteristics, and the results were similar to those shown in Figure 4.

4.3. Numerical Characterization. Figures 7 and 8 display the $R_{\max(S)}$ plots for the boiling point and Kazius data sets.

For the boiling point data set we consider those molecules that had less than 10% of the data set as neighbors at 50% of the maximum pairwise radius as outliers. This criterion identified three molecules, **226**, **249**, and **259**. We also considered a slightly looser criterion and raised the number of neighbor cutoff to 100. In this case, molecule **151** was added to the set of molecules noted above. One would thus expect that the data set is relatively well distributed through the given descriptor space. Figure 7 appears to confirm these observations. The *R*-NN curves for molecules **226** and **259** are shown in Figure 3. The maximum slope of the *R*-NN curves for these molecules corresponds to the linearly

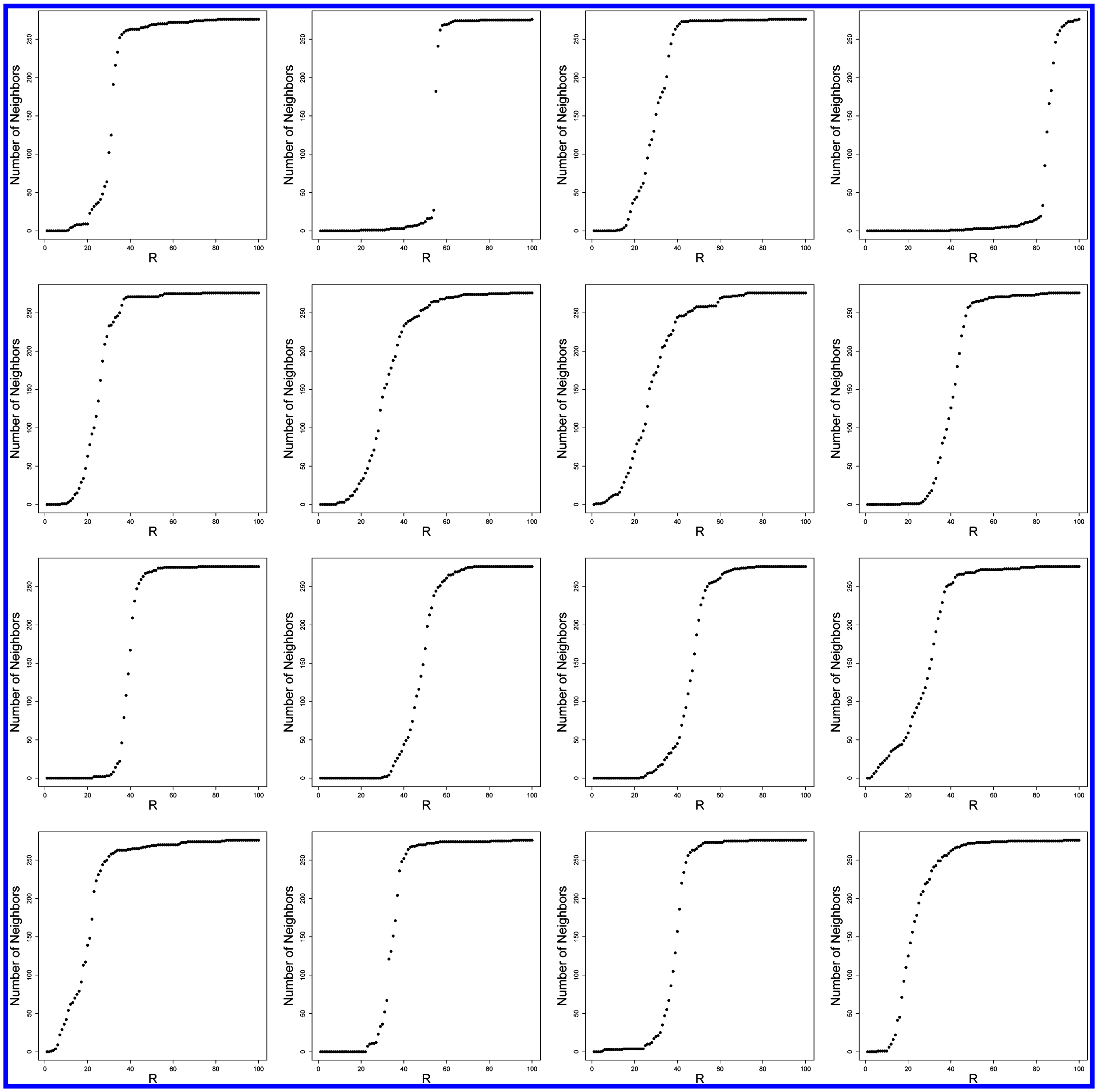
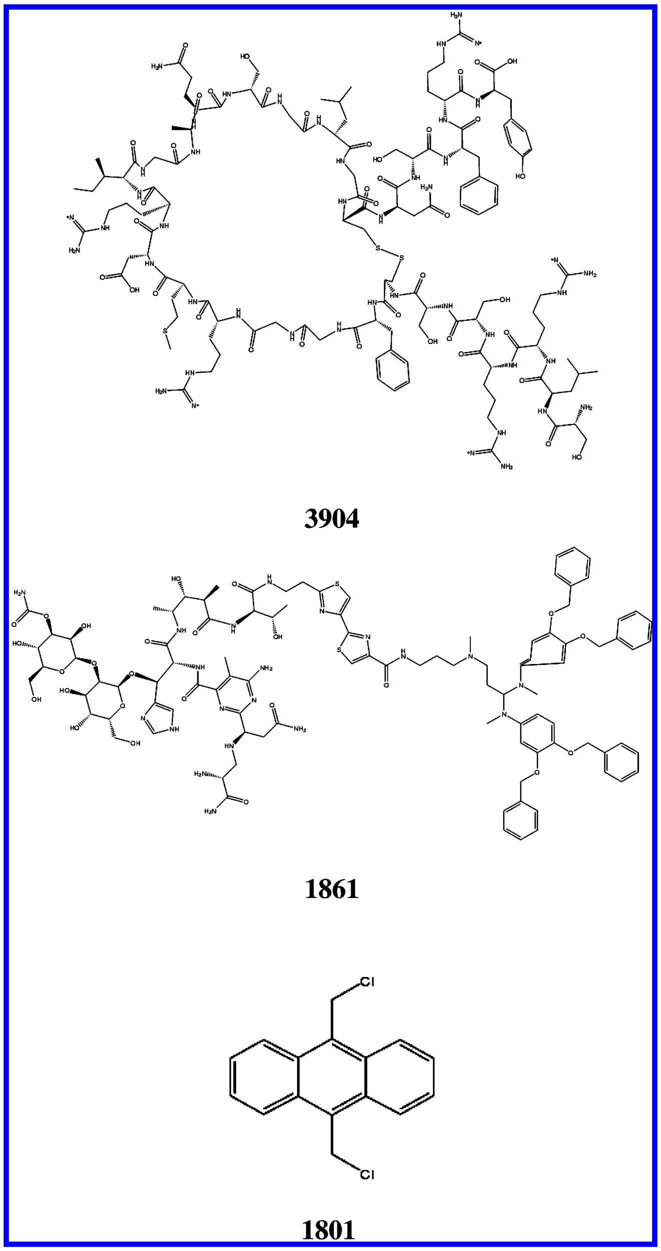


Figure 4. *R*-NN curves for molecule **226** from the boiling point data set obtained using 16 randomly selected descriptor subsets ranging in size from 3 to 8. Note that in all cases, the *R*-NN curve shows sigmoidal characteristics.

increasing portion of the sigmoidal curves. Clearly, if the radius for which this occurs is larger, it indicates that the molecule in question has very few neighbors (with respect to the size of the data set) for increasing radii and thus can be considered to occupy a sparse region of the descriptor space.

For the Kazius data set we followed a similar strategy as described above to first identify a set of outliers from the numerical data. That is, we considered molecules that had less than 10% of the data set as neighbors, at 50% of the maximum pairwise distance. This led to two molecules being flagged as outliers. The structures of these two outliers are shown in Figure 5. Since our constraints resulted in only two outliers, we considered molecules that had less than 10% of the data set as neighbors at a radius equal to 40% of the maximum pairwise distance. This resulted in six compounds, and Figure 9 displays the number of neighbors for these six compounds.

However this approach does require that the user define cutoff values for the radius and nearest neighbor count. Since these decisions are arbitrary, the use of the $R_{\max(S)}$ plot provides a simpler approach to easily identify outliers. Figure 8 shows the $R_{\max(S)}$ plot for the Kazius data set. The plot clearly distinguishes between these six outliers and the rest of the data set. It is also clear that even among the six outliers the ordering is maintained. That is, the most extreme outlier, compound **3904**, is isolated near the top of the plot. Though only the six outliers mentioned in the text have been annotated in Figure 8, it is clear that as we go down the *Y*-axis, we can create a ranking of the molecules in terms of their being located in sparse or dense regions of the data set. For this particular data set, visual inspection of the plot indicates that the majority of the molecules have $R_{\max(S)}$ less than 20, the average being approximately 8. The plot also indicates that a relatively small number of molecules have $R_{\max(S)}$ values that are significantly different from that of the



3904

1861

1801

Figure 5. Structures of the two most extreme outliers (**3904** and **1861**) from the Kazius data set. For comparison, molecule **1801** from a dense region of the descriptor space is also shown.

average value noted above. This appears to confirm our conjecture that the data set is quite evenly distributed in the descriptor space considered.

It is clear that using $R_{\max(S)}$ plots as opposed to a numerical characterization allows one to visually flag molecules in a data set as outliers. Furthermore, one is able to gradually classify the molecules ranging from extreme outliers to those that are not significant outliers but are not necessarily located in the bulk of the data set.

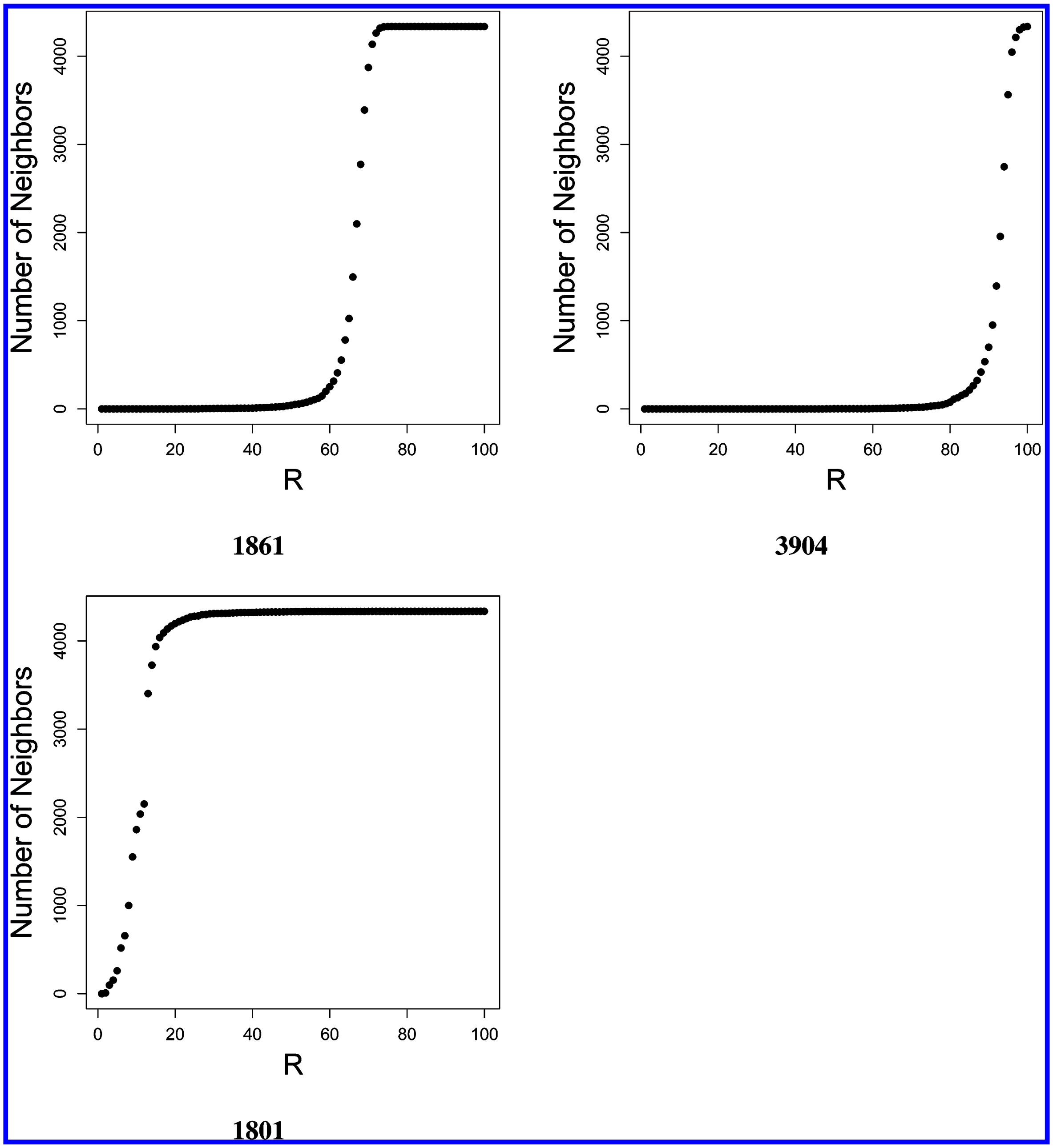
It should be noted that though we have defined $R_{\max(S)}$ to be the radius for which the slope is maximum, this can be changed. If one wanted to focus on the initial exponential portion of the sigmoidal curve, then one could consider slopes that are less than the maximum value, using a linear search for all values of the slope that have been calculated. Such an approach would allow one to focus in more detail on a subset of molecules if so desired. Our investigations indicate that considering $R_{\max(S)}$ as defined above is sufficient to highlight molecules lying in sparse regions of the data set.

5. DISCUSSION

The results described above have focused on analysis of individual data sets. More specifically we have not compared the number of neighbors of a given molecule with respect to different data sets. This can easily be achieved by normalizing the neighbor count by the total number of molecules in the data set. Such a transformation would allow sensible comparison of the location of a query molecule in different data sets. Another aspect of our approach is that it only considers single molecules. That is, the method indicates whether a molecule lies in a dense or sparse region of the descriptor space. It does not characterize the descriptor space as a whole in terms of sparsity. In this sense, *R*-NN curves can be considered a measure of local density for a given descriptor space. This concept is similar to that used by Daszykowski et al.[26] where they used the concept of local density to perform clustering. Future work involves the development of an approach to characterize the sparsity of a descriptor space as a whole as well as within given ranges of the component descriptors.

An important point to note about the technique presented here is that it makes no attempt to describe the suitability or utility of the specified descriptor space. That is, the technique only describes how molecules are distributed in a specific space. Thus, whether we analyze a descriptor space obtained from a *good* predictive model or randomly selected descriptors, the results would only describe the distribution of molecules in these spaces. It will not be able to say that a descriptor space is good or bad. This is an important feature as it allows the technique to be independent of the types of descriptors selected. At the same time the technique can be used to analyze the suitability of a given descriptor space. For example, if we consider the boiling point data set, we can see that molecule **226** is an outlier in some of the random descriptor spaces but not one in other spaces (Figure 4). Given multiple descriptor spaces it might be useful to be able to select a descriptor space where certain molecules are not outliers. This could be easily performed by obtaining the $R_{\max(S)}$ values for the molecules in different descriptor spaces and then plotting the values versus the index number of the descriptor space. However, for more than one molecule such a 2D plot would become unwieldy. Thus one could generate individual plots for each molecule which would only be feasible for a small number of molecules. Alternatively, one could create a 3D plot with the molecule index number, descriptor space index number, and $R_{\max(S)}$ values on the X, Y, and Z axes, respectively. A number of tools are available for the visualization[27] of such 3D plots, and we are currently investigating ways to perform this type of task in an intuitive fashion.

Our results confirm the theoretical analysis of nearest neighbor distributions. The algorithm to obtain *R*-NN curves is simple to implement, and the resultant visualization allows one to easily identify outliers in the data set. One concern of this method is the running time. Since the method essentially requires the calculation of pairwise distances for the data set, it has a time complexity of $O(n^2)$. For the boiling point data set the time required to generate the *R*-NN curve



1861

3904

1801

Figure 6. *R*-NN curves for three compounds from the Kazius data set. Molecules **1861** and **3904** are immediately identifiable as outliers. For comparison the *R*-NN curve for molecule **1801**, located in a dense region of the descriptor space, is shown.

data was less than 8 s (on a 2 GHz Pentium 4 with 512MB RAM, running Fedora Core 4). For the case of the Kazius data set the running time was 674 s. It is apparent that a brute force approach to the calculation of the *R*-NN curves for larger data sets is not feasible. However we have recently described[28] an approximate nearest neighbor algorithm, termed locality sensitive hashing (LSH), that avoids the calculation of the pairwise distance matrix and uses a geometric hashing scheme to obtain a set of approximate *R*-nearest neighbors with a specified probability. This algorithm has been shown to be 3−4 orders of magnitude faster than the traditional *k* NN algorithm and at least 94 accurate. Clearly, use of the LSH algorithm makes the calculation of *R*-NN curves a feasible task even for large data sets.

The approach to detection of outliers based on *R*-NN curves highlights an aspect of distance based methods such as *k* NN regression and classification which consider only the absolute nearest neighbors rather than the *R*-nearest neighbors. The fundamental assumption[29] made during QSAR modeling is that molecules that exhibit similar structural features (as characterized by molecular descriptors) will exhibit similar values of the measured property. Of course this cannot be taken as a firm rule as shown by Martin et al.[29] and as discussed by Kubinyi.[30] However this assumption is used directly in *k* NN regression and classification where the predicted property of a query molecule is taken to be the average of the observed properties of the *k* nearest neighbors. Our analysis has shown that in a number of cases, a query point is located far from the bulk of the data set, and as a result the *nearest* molecules are really quite far away. Thus for such query molecules, the fundamental assumption noted above does not necessarily hold. One would expect that the predicted value (or class) of such

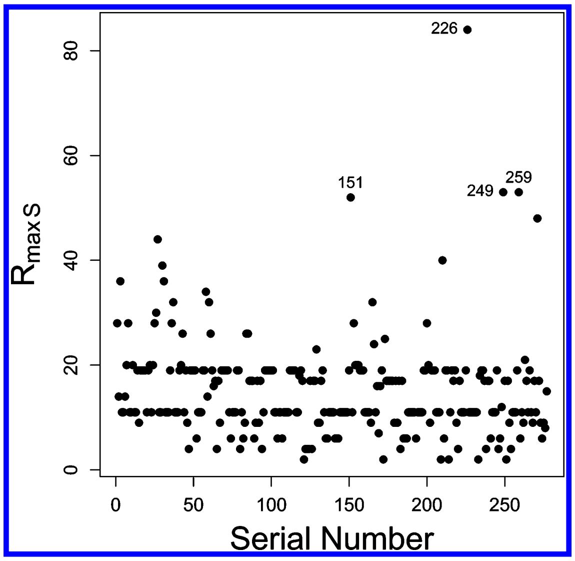


Figure 7. A plot of $R_{\max(S)}$, the radius for which the slope of the curve is maximum for the boiling point data set. The annotated points have less than 100 neighbors at a radius equal to 50 of the maximum pairwise distance and thus are regarded as outliers.

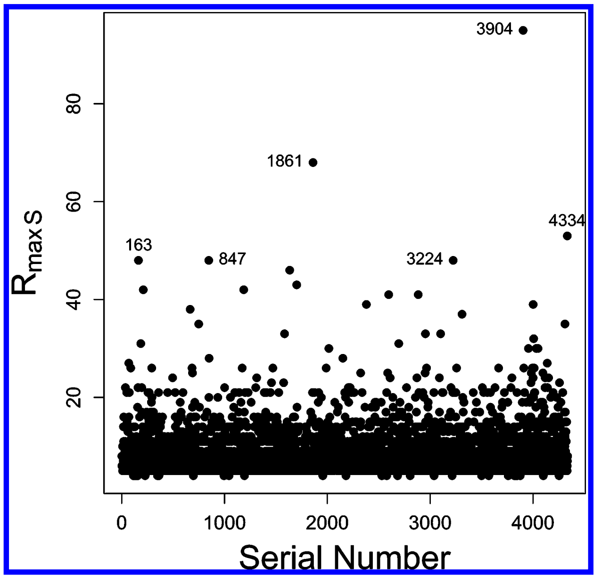


Figure 8. A plot of $R_{\max(S)}$, the radius for which the slope of the curve is maximum for the Kazius data set. The uppermost annotated points have less than 100 neighbors at a radius equal to 50% of the maximum pairwise distance and thus are regarded as extreme outliers. All other annotated points had less than 433 neighbors at 40% of the maximum pairwise distance.

molecules would not be very reliable. Given this observation, we believe that the traditional use of the *k* NN method[31−34] can lead to misleading results for certain query molecules. To avoid this we feel that an *R*-NN analysis of the data set and query molecules should be performed to ascertain that the query molecule does indeed lie in a relatively dense region of the descriptor space so that the query molecule will have molecules close to it.

Given the above discussion of the use of *R*-NN curves for the purposes of outlier detection we should also point out that another method for identifying outliers visually is to use principal component (PC) plots, where one would plot the first versus second principal components for a data set. Most distinct outliers will be located away from the bulk of the plot (assuming absence of clustering). We generated principal components plots (using only the first and second principal components), and the extreme outliers detected by the *R*-NN curve approach were also observed as outliers on the PC plots. Though the PC plot approach to outlier detection is simple in practice, it has a number of disadvantages. First, the PC plot is a suitable tool for visualizing the data set and the subsequent visual identification of outliers, but it does not, by itself, yield an algorithm to identify the outliers automatically. That is, PC analysis merely reduces the dimension of the data set along the most relevant low dimensional components. We would still need to run some other algorithm to detect outliers automatically, and our *R*-nn curve approach could be one of those methods, though, as has been shown, one can also use the *R*-NN curve technique without performing dimension reduction beforehand. In addition a principal components analysis involves an eigendecomposition of the data set. This is usually performed using the Singular Value Decomposition (SVD), which has a time complexity of $O(\min(mn^2, nm^2))$, where m is the number of dimensions of the data set and n is the number of rows in the data set. Thus for very large data sets of high dimensionality, the SVD can become computationally infeasible. Though approximate SVD algorithms are available,[35] their use can lead to inaccuracies in the final principal components. On the other hand, such data sets can be analyzed using the LSH algorithm which has been shown to run in subquadratic time in the number of points (or rows) and linear in the number of dimensions.[28] Second, the use of PC plots generally limits one to viewing pairs of principal components. For low dimensional data sets (3−4 descriptors) this is not a significant problem. However for larger dimensions it is possible that an outlying molecule is not identified as such on the plot of the first two PCs. Instead it may be identified as an outlier for some other pair of PCs. Clearly, the use of PC plots alone to obtain a global view of data set sparsity can become unwieldy and even misleading. This is especially true if the characteristic of the high dimensional chemical space is nonlinear. In such cases we might first have to project the nonlinear space on a low

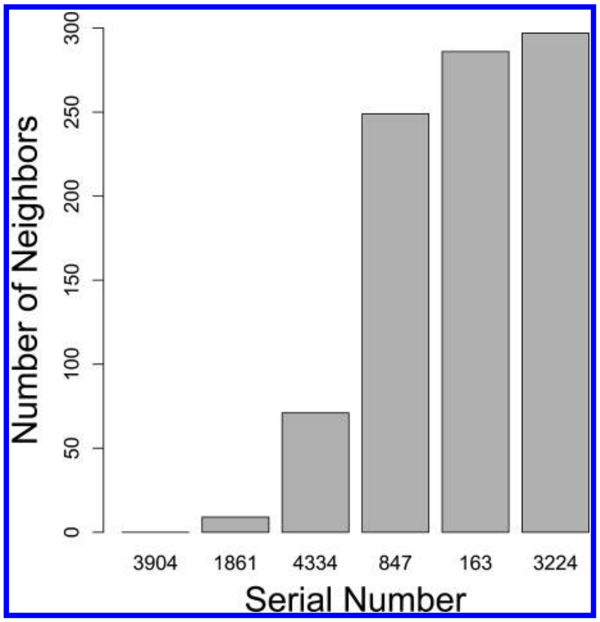


Figure 9. The number of neighbors for the molecules from the Kazius data set identified as outliers at a radius equal to 40% of the maximum pairwise distance.

dimension manifold to help us visually detect those outliers. On the other hand, since the *R*-NN curve approach utilizes all the dimensions of the data set, the plots provide an easily understandable view of the molecules location in descriptor space. Furthermore, the numerical characterization of the curves allows one to easily analyze the data set as a whole rather than considering specific molecules.

6. CONCLUSIONS

We have presented a simple and intuitive approach to the problem of outlier detection for data sets of arbitrary dimensions. The method is based on counts of neighbors lying within a radius *R* of a query molecule for varying *R*. The method results in a sigmoidal curve for a query molecule which provides easy visual characterization of the location of the molecule in a given descriptor space. For molecules lying in dense and sparse regions of the descriptor space the curves are easily differentiated by the lower tail of the sigmoidal curve. However visual inspection of the curves for a large collection of molecules can be unwieldy, and hence we provide a numerical characterization of the curves allowing a summary of the *outlyingness* of the molecules in a data set in a single graph. Furthermore, given precalculated *R*-NN curves for a data set, one can easily vary the constraints (number of neighbors and radius) that define an outlier, thus allowing one to focus on molecules that occupy increasingly sparse (or dense) regions of the descriptor space.

REFERENCES AND NOTES

(1) Pearlman, R.; Smith, K. Metric Validation And The Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(2) Pearlman, R.; Smith, K. Novel Software Tools For Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, 339−353.

(3) Schnur, D. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 36−45.

(4) Cramer, R. D.; Jilek, R. J.; Guessregen, S.; Clark, S. J.; Wendt, B.; Clark, R. D. "Lead Hopping". Validation of Topomer Similarity as a Superior Predictor of Similar Biological Activities. *J. Med. Chem.* **2004**, *47*, 6777−6791.

(5) Saeh, J. C.; Lyne, P. D.; Takasaki, B. K.; Cosgrove, D. A. Lead Hopping Using SVM and 3D Pharmacophore Fingerprints. *J. Chem. Inf. Comput. Sci.* **2005**, *45*, 1122−1133.

(6) Jorgensen, W. Design and Diversity Analysis of Large Combinatorial Libraries Using Cell-Based Methods. *Science* **2004**, *303*, 1813−1818.

(7) Agrafiotis, D. K.; Lobanov, V. S. An Efficient Implementation of Distance-Based Diversity Measures Based on *k*-d Trees. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 51−58.

(8) Agrafiotis, D. K. A Constant Time Algorithm for Estimating the Diversity of Large Chemical Libraries. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 156−167.

(9) Godden, J. W.; Xue, L.; Kitchen, D. B.; Stahura, F. L.; Schermerhorn, E. J.; Bajorath, J. Median Partitioning: A Novel Method for the Selection of Representative Subsets from Large Compound Pools. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 885−893.

(10) Makara, G. Measuring Molecular Similarity and Diversity: Total Pharmacophore Diversity. *J. Med. Chem.* **2001**, *44*, 3563−3571.

(11) Trepalin, S. V.; Gerasimenko, V. A.; Kozyukov, A. V.; Savchuk, N. P.; Ivaschenko, A. A. New Diversity Calculations Algorithms Used for Compound Selection. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 249−258.

(12) Cummins, D. J.; Andrews, C. W.; Bentley, J. A.; Cory, M. Molecular Diversity in Chemical Databases: Comparison of Medicinal Chemistry Knowledge Bases and Databases of Commercially Available Compounds. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 750−763.

(13) Pearlman, R. S.; Smith, K. M. Metric Validation and the Receptor-Relevant Subspace Concept. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 28−35.

(14) Golbraikh, A. Molecular data set Diversity Indices and Their Applications to Comparison of Chemical Databases and QSAR Analysis. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 414−425.

(15) Pickett, S.; Mason, J.; McLay, I. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.

(16) Mount, J.; Rupert, J.; Welch, W.; Jain, A. IcePick: A Flexible Surface Based System for Molecular Diversity. *J. Med. Chem.* **1999**, *42*, 60−66.

(17) Agrafiotis, D. On the Use of Information Theory for Assessing Molecular Diversity. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 576−580.

(18) Agrafiotis, D.; Rassokhin, D. A Fractal Approach for Selecting an Appropriate Bin Size for Cell-Based Diversity Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 117−122.

(19) Datar, M.; Immorlica, N.; Indyk, P.; Mirrokni, V. S. Locality-Sensitive Hashing Scheme Based on p-Stable Distributions. In *SCG '04: Proceedings of the twentieth annual symposium on Computational geometry*; ACM Press: New York, U.S.A., 2004.

(20) Zhang, M.; Vassiliadis, S.; Delgado-Frias, J. G. Sigmoid Generators for Neural Computing Using Piecewise Approximations. *IEEE Trans. Comput.* **1996**, *45*, 1045−1049.

(21) Goll, E.; Jurs, P. Prediction of the Normal Boiling Points of Organic Compounds From Molecular Structures with a Computational Neural Network Model. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 974−983.

(22) Chemical Computing Group Inc. *Molecular Operating Environment (MOE 2004.03)*.

(23) Kazius, J.; McGuire, R.; Bursi, R. Derivation and Validation of Toxicophores for Mutagenicity Prediction. *J. Med. Chem.* **2005**, *48*, 312−320.

(24) 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.

(25) Shanmugasundaram, V.; Maggiora, G.; Lajiness, M. Hit-Directed Nearest Neighbor Searching. *J. Med. Chem.* **2005**, *48*, 240−248.

(26) Daszykowski, M.; Walczak, B.; Massart, D. L. Looking for Natural Patterns in Analytical Data. 2. Tracing Local Density with OPTICS. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 500−507.

(27) Swayne, D.; Buja, A.; Temple-Lang, D. Exploratory Visual Analysis of Graphs in GGobi. In *Proceedings of the 3rd International Workshop on Dist. Stat. Comp.*; 2003.

(28) 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.

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

(30) Kubinyi, H. Similarity and Dissimilarity − A Medicinal Chemists Perspective. *Perspect. Drug Discovery Des.* **1998**, *11*, 225−252.

(31) Ajmani, S.; Jadhav, K.; Kulkarni, S. A. Three-Dimensional QSAR Using the *k*-Nearest Neighbor Method and Its Interpretation. *J. Chem. Inf. Model.* **2005**, ASAP.

(32) Itskowitz, P.; Tropsha, A. *k*-Nearest Neighbors QSAR Modeling as a Variational Problem: Theory and Applications. *J. Chem. Inf. Model.* **2005**, *45*, 777−785.

(33) Braga, S.; Galvão, D. Benzo[*c*]quinolizin-3-ones Theoretical Investigation: SAR Analysis and Application to Nontested Compounds. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1987−1997.

(34) Souza, J., J.; Molfetta, F. A.; Honorio, K. M.; Santos, R. H. A.; da Silva, A. B. F. A Study on the Antipicornavirus Activity of Flavonoid Compounds (Flavones) by Using Quantum Chemical and Chemometric Methods. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1153−1161.

(35) Drineas, P.; Frieze, A.; Kannan, R.; Vempala, S.; Vinay, V. Clustering Large Graphs via the Singular Value Decomposition. *Mach. Learn.* **2004**, *56*, 9−33.

CI060013H