Structure−Activity Landscape Index: Identifying and Quantifying Activity Cliffs

Rajarshi Guha[‡] and John H. Van Drie*[,†]

School of Informatics, Indiana University, Bloomington, Indiana 47406, and John H. Van Drie Research LLC,
Andover, Massachusetts 01810

Received November 8, 2007

A new method for analyzing a structure−activity relationship is proposed. By use of a simple quantitative index, one can readily identify "structure−activity cliffs": pairs of molecules which are most similar but have the largest change in activity. We show how this provides a graphical representation of the entire SAR, in a way that allows the salient features of the SAR to be quickly grasped. In addition, the approach allows us view the SARs in a data set at different levels of detail. The method is tested on two data sets that highlight its ability to easily extract SAR information. Finally, we demonstrate that this method is robust using a variety of computational control experiments and discuss possible applications of this technique to QSAR model evaluation.

1. INTRODUCTION

In a recent article entitled "On outliers and activity cliffs - why QSAR often disappoints", G. Maggiora[1] proposes that QSAR models tend to disappoint because the underlying structure−activity landscape resembles more the sharp cliffs of Bryce Canyon, while QSAR makes an implicit assumption that the landscape looks like the smooth hills of Kansas.

We follow up on this idea here, in a more precise way, by showing how one may analyze an SAR pairwise, in a way that highlights the sharply varying regions, i.e., those pairs of molecules which are very similar but have large changes in activity. This pairwise analysis leads to structure−activity cliffs, and the whole set of cliffs leads to a graph representation of an SAR where each node is a molecule, and each edge is a structure activity cliff. This pairwise view more closely mimics how medicinal chemists view an SAR. These pairs arise from the calculation of an index, the SALI (structure−activity landscape index), defined below and is based on developments in the mid-1990s.[2] This analysis is only appropriate for biological activities that are receptor-mediated, i.e., it should not be used to analyze physico-chemical properties such as solubility.

A number of workers have addressed the problem of analyzing compounds in a pairwise fashion to identify how structural differences correlate with changes in activity. Examples of approaches include the technique of "matched molecular pairs" described by Leach et al.[3] and SAR maps described by Agrafiotis et al.[4] It should be noted that both of these approaches are focused on the visualization of structures in such a way as to help examine large data sets for SARs. Recently Peltason et al.[5] described the SAR Index (SARI) which aims to quantitatively characterize SARs present in data based on 2D fingerprints and a reference compound. The method was subsequently used to characterize SARs into various classes. In comparison to these approaches, our methodology is more general in that it only depends on the structures in the data set. Though the method is not completely automated, the single user definable parameter provides the user with the flexibility of identifying different types of SARs. Furthermore, our approach allows both the visualiation of SARs present in a data set as well as utilizing these trends to assess model quality for arbitrary models.

To ensure that our analyses are objective, we submitted them to a number of computational control experiments, e.g., applying it to a variety of data sets, varying the similarity metrics used, and testing the effect of noise added to the biological data.

2. METHODOLOGY

Given the concept of structure−activity cliffs, the first step is to quantify this concept. Our approach involves defining a structure−activity landscape index (SALI) as

$$\mathrm{SALI}_{i,j} = \frac{|A_i - A_j|}{1 - \mathrm{sim}(i,j)} \tag{1}$$

where A_i and A_j are the activities of the *i*th and the *j*th molecules, and sim(*i*,*j*) is the similarity coefficient between the two molecules. It is clear that the most significant activity cliffs in a data set will lead to high SALI values, i.e., structurally similar compounds with large changes in activity; in general it is these values that will be of interest to the scientist.

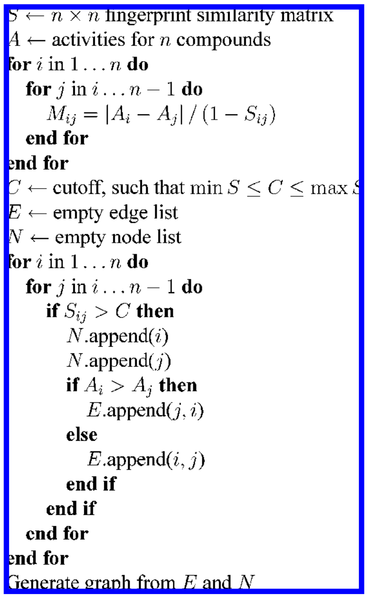
It is possible that two very similar, but not identical, compounds will have a similarity value of 1 (e.g., stereoisomers). In such a case the SALI value will be infinity. In our calculations we have replaced such values with the next largest SALI value. In this study we calculate similarity using the Tanimoto coefficient evaluated between pairs of fingerprints; it will be shown that the overall analyses depends minimally on the specific similarity metric and fingerprint type.

One typically uses either the raw activity value or the log of that value. The former is more appropriate for activity

* Corresponding author e-mail: johnvandrie@mindspring.com.
[†] John H. Van Drie Research LLC.
[‡] Indiana University.

10.1021/ci7004093 CCC: $40.75 © 2008 American Chemical Society
Published on Web 02/28/2008

Chart 1. Procedure To Generate a SALI Graph from Binary Fingerprint and Biological Activity Data



```
S ← $n \times n$ fingerprint similarity matrix
A ← activities for $n$ compounds
for $i$ in $1 \ldots n$ do
    for $j$ in $i \ldots n-1$ do
        $M_{ij} = |A_i - A_j| \,/\, (1 - S_{ij})$
    end for
end for
C ← cutoff, such that $\min S \leq C \leq \max S$
E ← empty edge list
N ← empty node list
for $i$ in $1 \ldots n$ do
    for $j$ in $i \ldots n-1$ do
        if $S_{ij} > C$ then
            $N$.append($i$)
            $N$.append($j$)
            if $A_i > A_j$ then
                $E$.append($j, i$)
            else
                $E$.append($i, j$)
            end if
        end if
    end for
end for
Generate graph from $E$ and $N$
```

which is represented as percent inhibition; the latter, when the activity is a K_i or IC_{50}.

Given a set of N structures, we obtain an $N \times N$ matrix of SALI values. The simplest approach is to simply sort that list by their SALI value, bringing to the top those pairs of structures most similar with the largest change in activity. This approach yields a purely "local" view of the SAR.

To gain a more "global" view of the SAR, we considered a more general approach where the SALI values were analyzed to generate a graph representation of structure−activity cliffs. In this representation, each compound is a node, and two nodes are connected if their SALI value is greater than a user specified cutoff. The procedure used to generate the data for the graph representation from the SALI matrix is shown in Chart 1. The cutoff is bounded by the maximum and minimum values of the SALI matrix generated for the data set. The goal of the cutoff is to allow the user to focus on increasingly significant activity cliffs. Thus, for a low value of the cutoff, the bulk of the compounds will be connected, leading to a very dense graph representation. As the cutoff is increased, fewer pairs of compounds remain connected, and as the cutoff tends toward the maximum SALI value, only the most significant activity cliffs remain. An example of such a SALI graph is shown in Figure 1. In this graph, the edges are directed such that the node at the tail of an edge has a lower activity than a node at the head of an edge. However, for a given graph, there is no special ordering for a given horizontal level.

Admittedly, it is tricky to view these graphs in a static printed page. An interactive software tool is the ideal way to navigate such graphs. We also note that though the cutoff is bounded, the actual value is user-defined and depends on the level of detail one wishes the network to have. Since this depends on the nature of the data set, it is difficult to specify a general cutoff value. Furthermore, once the SALI matrix has been calculated, the network can be regenerated very rapidly, for different values of the cutoff.

Another possible method to gain a global view of the SAR is to plot a "heatmap" image of the values. An example of this method is shown in Figure 2. The X and Y axes represent the compound indices and are sorted in increasing order of activity values. White represents pairs of compounds exhibiting the highest SALI values and black indicates pairs exhibiting the minimum SALI values. Given the ordering of compounds by activity, we immediately note two types of compounds.

Lighter blocks toward the upper right indicate compounds that are both highly active and still represent an activity cliff. This implies that such pairs represent successful efforts toward optimizing lead compounds. On the other hand, light blocks located in the upper left corner indicate pairs of compounds, of which one member is active and the other inactive, exhibiting an activity cliff. Such compound pairs represent efforts that were able to identify lead compounds, rather than efforts focused on optimization.

Virtually any attempt to model structure−activity data relies implicitly on the assumption that the measured biological data results from a specific molecular recognition event. However, it is not an infrequent occurrence that one encounters a "flat SAR", where the biological data changes little over wide variations in chemical structure. Frequently, a "flat SAR" occurs when the biological response arises from an artifactual mechanism, e.g., membrane perturbation, or aggregation. When a "flat SAR" is encountered, these SALI heatmaps allow one to easily identify this characteristic, alerting modelers to be wary of applying typical modeling techniques, and alerting medicinal chemists to the possibility that their SAR is not real.

For all data sets we evaluated 1052 bit BCI structural fingerprints[6] and performed the analysis and visualization using R[7] and the Rgraphviz package[8] which interfaces the Graphviz layout tool[9] to the R environment. Graphviz provides a number of layout algorithms, and the SALI networks shown in this paper were generated using the "dot" method. All calculations were performed on a MacBook Pro (Intel dual core 2.2 GHz) with 1GB RAM and running MacOS X 10.4. The R source code to evaluate SALI matrices and generate the static network visualizations are provided as Supporting Information. An interactive application based on the ZGRViewer[10] toolkit is available from http://cheminfo.informatics.indiana.edu/~rguha/code/java/salivis. The application allows one to visualize and explore precalculated SALI networks as well as generating networks on the fly, from SMILES and activity data.

3. APPLICATION TO RELEVANT DATA SETS

A number of data sets were used for this study. The first data set consisted of 79 derivatives of 4-piperazinylquinazolines studied for their ability to inhibit PDGFR.[11] We primarily used this data set to calibrate the properties of the SALI values.

The second data set was a collection of 81 molecules that were studied as agonists[12] and antagonists[13] of the human melanocortin-4 receptor. The third data set consisted of 62 derivatives[14,15] that were designed to inhibit the glucocorticoid receptor. We specifically chose the last two data sets for this study as the biological activity reported for a number of compound pairs exhibited significant activity cliffs, which our method should be able to highlight.

3.1. Melanocortin-4 Data Set. This data set consists of 81 pyrrolidines and pyrrolidinones that were studied as

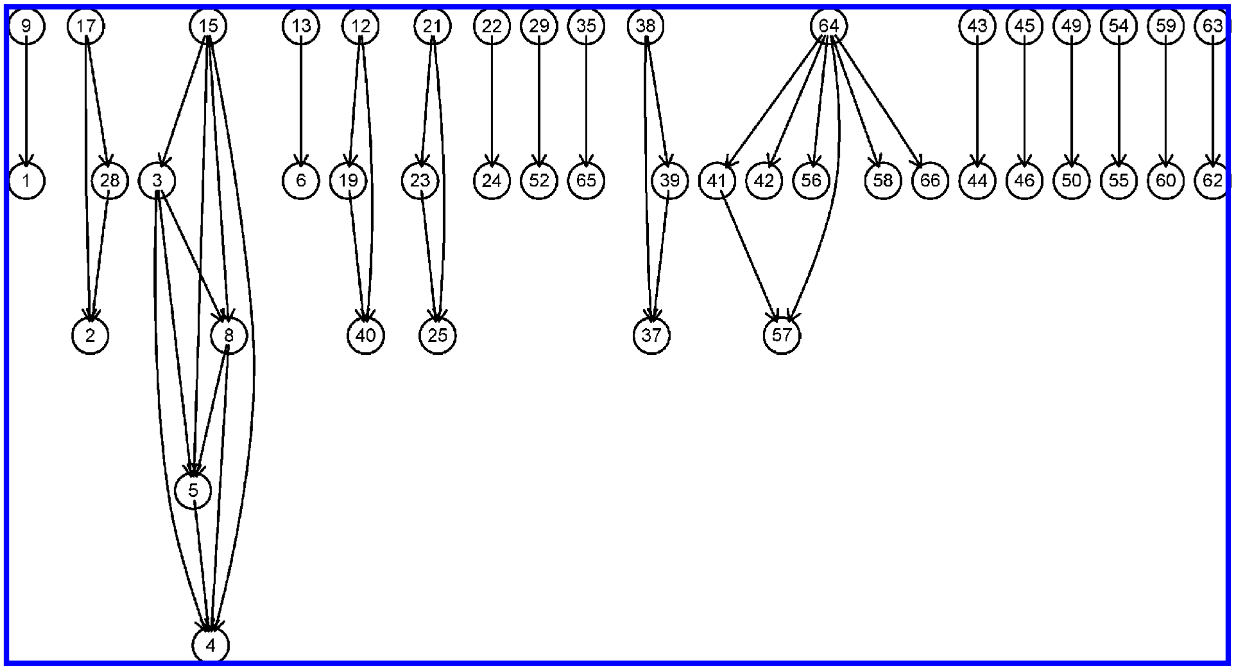


Figure 1. A graph representation of the SALI values. Numbers identify molecules. An edge occurs between two nodes if the SALI value for that pair is greater than a user-specified cutoff. The edges are directed such that the head node corresponds to a molecule with higher activity than the molecule corresponding to the tail node.

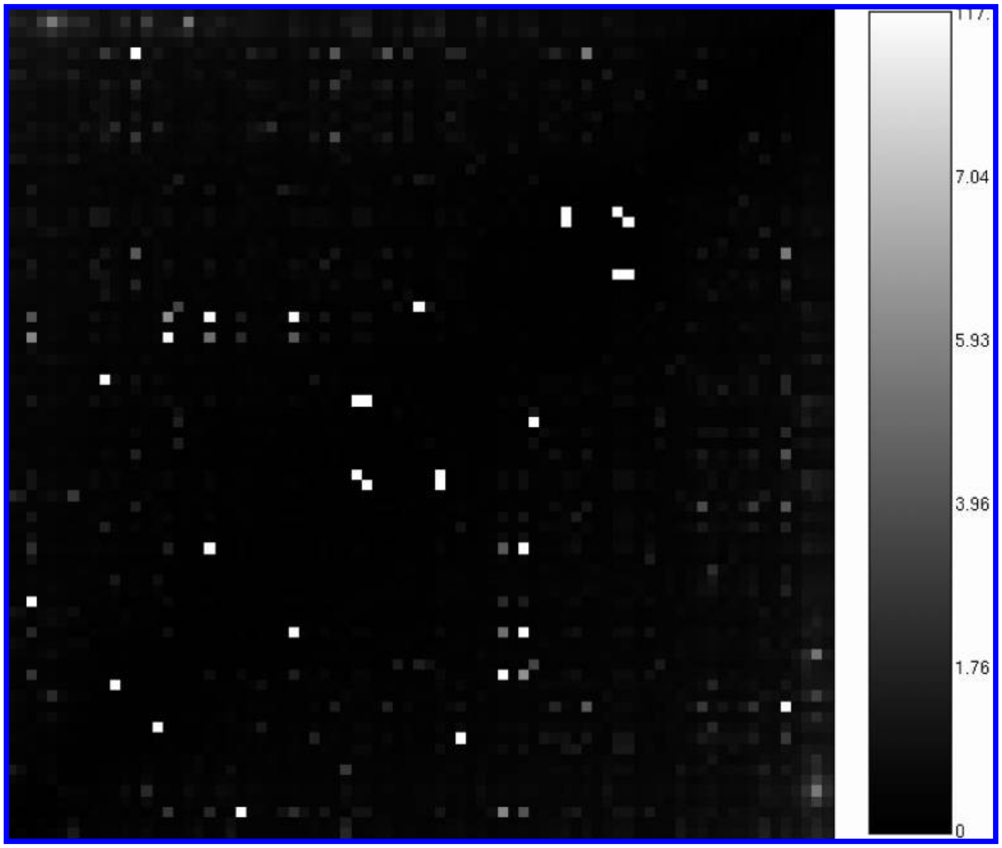
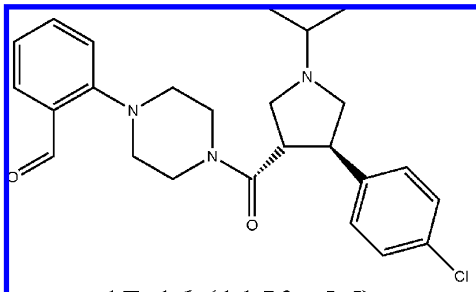


Figure 2. A heatmap representation of SALI values for the PDGFR data set using 1052 bit BCI fingerprints and the Tanimoto metric. The X and Y axes correspond to compounds and are ordered in terms of increasing activity. The legend indicates the range of SALI values.

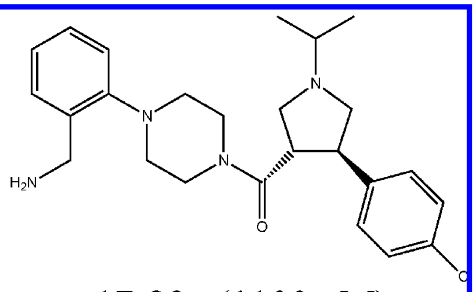
agonists[12] and antagonists[13] of the human melanocortin-4 receptor. These molecules were composed of 27 antagonists and 54 agonists. Given their opposing behavior as well as representing distinct synthetic series, it was expected that the collection of 81 molecules would exhibit multiple structure−activity relationships, and we were curious to see if this emerged in the SALI analysis. Only the K_i (nM) values were considered for the study.

The structures for these compounds are displayed in Table 1. For this data set we label molecules in the form *volume-number* where *volume* is the volume number of the publication that the molecule was taken from and *number* is the

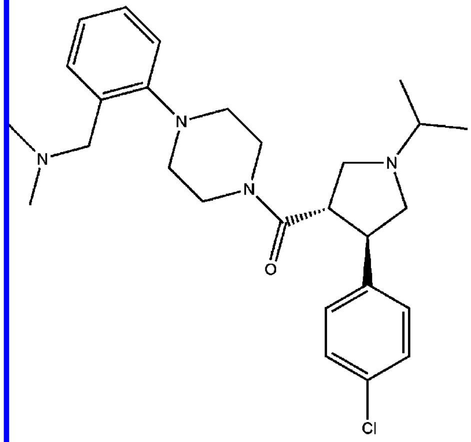
Table 1. Melanocortin-4 Receptor Inhibitors[a]



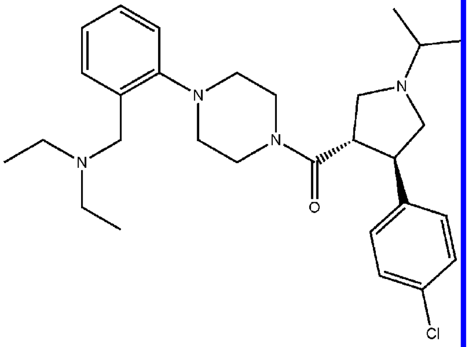
17-16 (1150 nM)



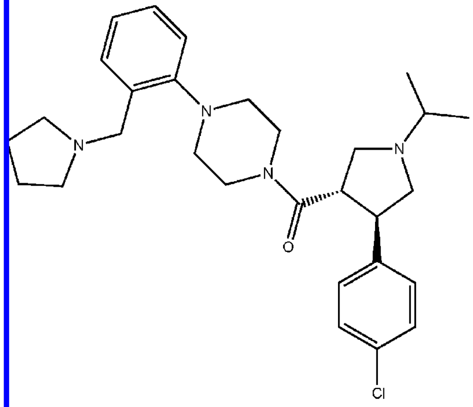
17-20a (1100 nM)



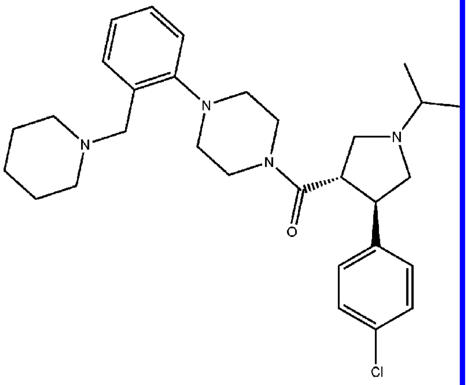
17-23a (1200 nM)



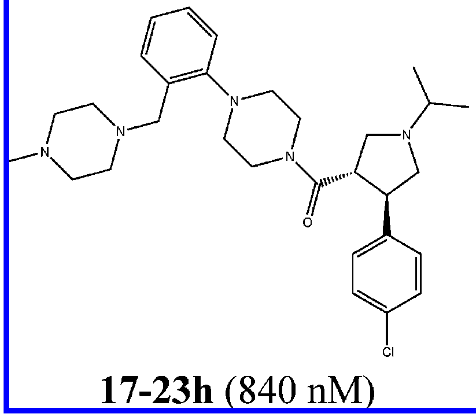
17-23b (1200 nM)



17-23e (550 nM)



17-23f (950 nM)



17-23h (840 nM)

[a] The values in parentheses indicate the experimentally determined K_i for the compound.

number used by the authors to refer to that molecule in the original publication.

The SALI networks using two different values of the cutoffs are shown in Figure 3. It should be noted that we

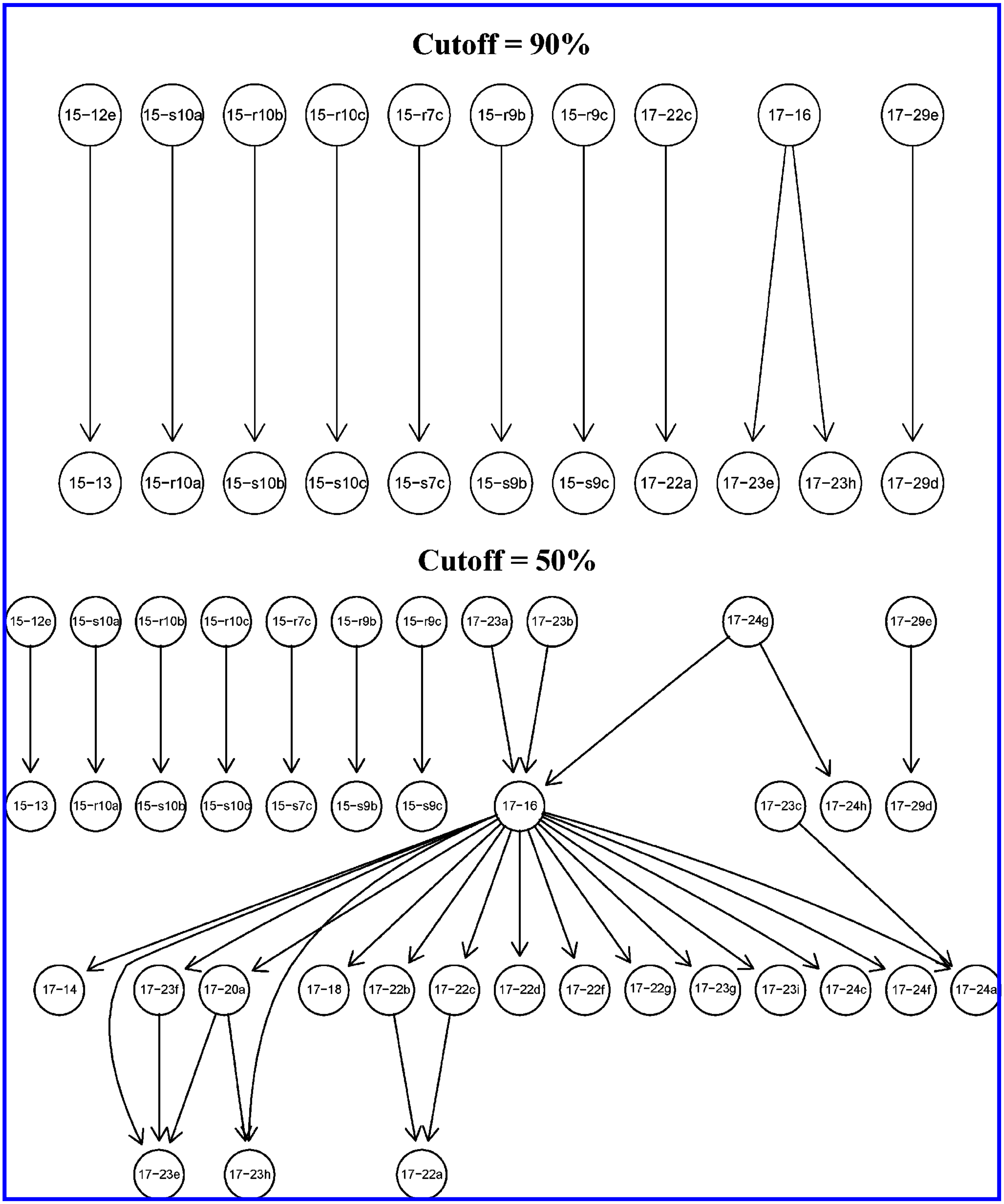
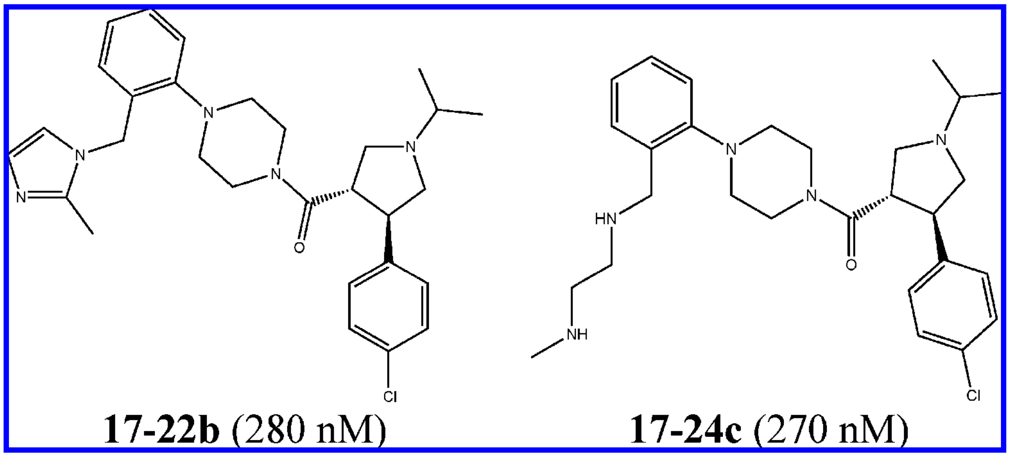


Figure 3. SALI networks for the melanocortin-4 receptor inhibitor data set. The cutoffs are reported as a percentage of the range of SALI values for the data set.

excluded **17-15** while generating the graphs. This was due to the fact that the K_i value for this molecule was significantly larger than for the rest of the data set, and so the resultant graphs did not show any significant structure. Though quite arbitrary, this is necessay to avoid compounds, whose measured activity may be a censored value or in error, from overshadowing other compounds. We discuss the issue of experimental outliers in more detail in section 5.

The graph generated using a 90% cutoff is quite simple, indicating that only the most significant structural changes that lead to an increase in potency are highlighted. We see that the graph consists primarily of disconnected edges. Most

Table 2. Melanocortin−4 Receptor Inhibitors That Exhibit Similar Activities, yet Are Structurally Similar[a]



17-22b (280 nM) **17-24c** (270 nM)

[a] The values in parentheses indicate the experimentally determined K_i for the compound.

of these edges correspond to enantiomeric pairs (such as **15-r10b** and **15-s10b**). Given that the fingerprints employed in this study do not capture stereochemistry, such pairs will have a similarity value of 1.0 and hence a SALI value of infinity. If we then consider **17-16** we see that it is connected to two other compounds, **17-23e** and **17-23h**. From Table 1 we see that the difference in activity for the two pairs is very large and is associated with an aldehyde group being changed to a pyrrolidine (in **17-23e**) or a 4-methylpiperazine (in **17-23h**) group. It would appear that the SAR represented by these two edges indicates that bulky groups at the benzylic position will lead to an increase in potency. Furthermore, the transition from an aldehyde to amine groups also appears to play a role in increasing potency, though simply substituting an amino group for an aldehyde group (**17-16** vs **17-20a**) does not lead to a significant increase in potency. These observations are in line with those made by the authors where they note that addition of amino groups at the benzylic position generally led improved potency. At the same time, if we consider the 50% graph we see that compounds **17-23a** and **17-23b** are both marginally less active than **17-16**, but both have an amine group at the benzylic position. It is clear that in these two molecules, the benzylic group is bulkier than the aldehyde and also contains an amino group. This would seem to suggest that the SAR not only is simply based on bulk but also might consider electronic effects such as H-bond donor/acceptor behavior.

The fact that the SAR is not necessarily a linear function of bulk is exemplified if we consider the 50% graph and the path **17-16** → **17-23f** → **17-23e**. We see that the aldehyde group is replaced by a piperidine (changing potency from 1150 nM to 950 nM), which when replaced by a pyrrolidine group further improves the potency to 550 nM. That is, even though the piperidine group is bulkier than the pyrrolidine, this does not lead to maximal potency. A similar trend can be seen if we note the paths **17-16** → **17-20a** → **17-23e** and **17-16** → **17-20a** → **17-23h**. In this case we see that going from **17-16** to **17-20a** improves potency due to the benzylic amine. However the increase is not very large, presumably due to the lack of bulk at this position. However going from **17-20a** to **17-23e** or **17-23h** does lead to a jump in potency, by virtue of an increasingly bulky group, but the increase does not correlate directly with the size of the group.

The above discussion focuses entirely on the activity cliffs. The observations can be strengthened by also considering pairs of molecules that have values below the cutoff. The current network representation does not allow us to easily identify such pairs, as the graph becomes very complex (since by definition, the bulk of the data set will lie below the cutoff). Instead, it easier to navigate the heatmap representation of the SALI data to identify such pairs for a given molecule. For example, if we compare compounds **17-22b** and **17-24c** (Table 2), this pair of compounds has very different substitutions at the benzylic position—a methylimidazole versus an aliphatic chain—yet has very similar activities. Though both these compounds exist as nodes in the 50% graph, they are not connected, since their SALI value does not satisfy the cutoff. The fact that the molecules are structurally different, yet have similar activities, is suggestive of the fact that the substituents act as bioisosteres.[16] At the same time, this pairs support the SAR described above, since both groups are relatively bulky and the compounds have a relatively low K_i value.

These observations highlight an important aspect of a SALI network. Namely, they can be used to identify what we term “large” and “small” SARs. Thus if one considers the 90% graph we see that by going from an aldehyde to a large nitrogen containing group increases potency. This is a very simple trend, but the interesting details are hidden. Is the increase in potency due to the presence of the nitrogen groups? Or is it due to the bulk? To answer these questions we must consider the more detailed 50% graph. As noted above, we can now break down the very broad SAR identified from the 90% graph and easily observe that an amino group at the benzylic position is indeed required to increase potency, but on its own it does not lead to a significant increase. To further improve potency we are required to increase the bulk at that position. But a continuous increase in bulk does not necessarily lead to a continuous increase in potency. Thus, by increasing the resolution of the network we are able to dissect a broad SAR into a set of more detailed components.

3.2. Glucocorticoid Data Set. The third data set consisted of 62 dihydroquinoline derivatives[14,15] that were designed to inhibit the glucocorticoid receptor. For this data set, the authors reported IC_{50} (nM) values against the glucocorticoid

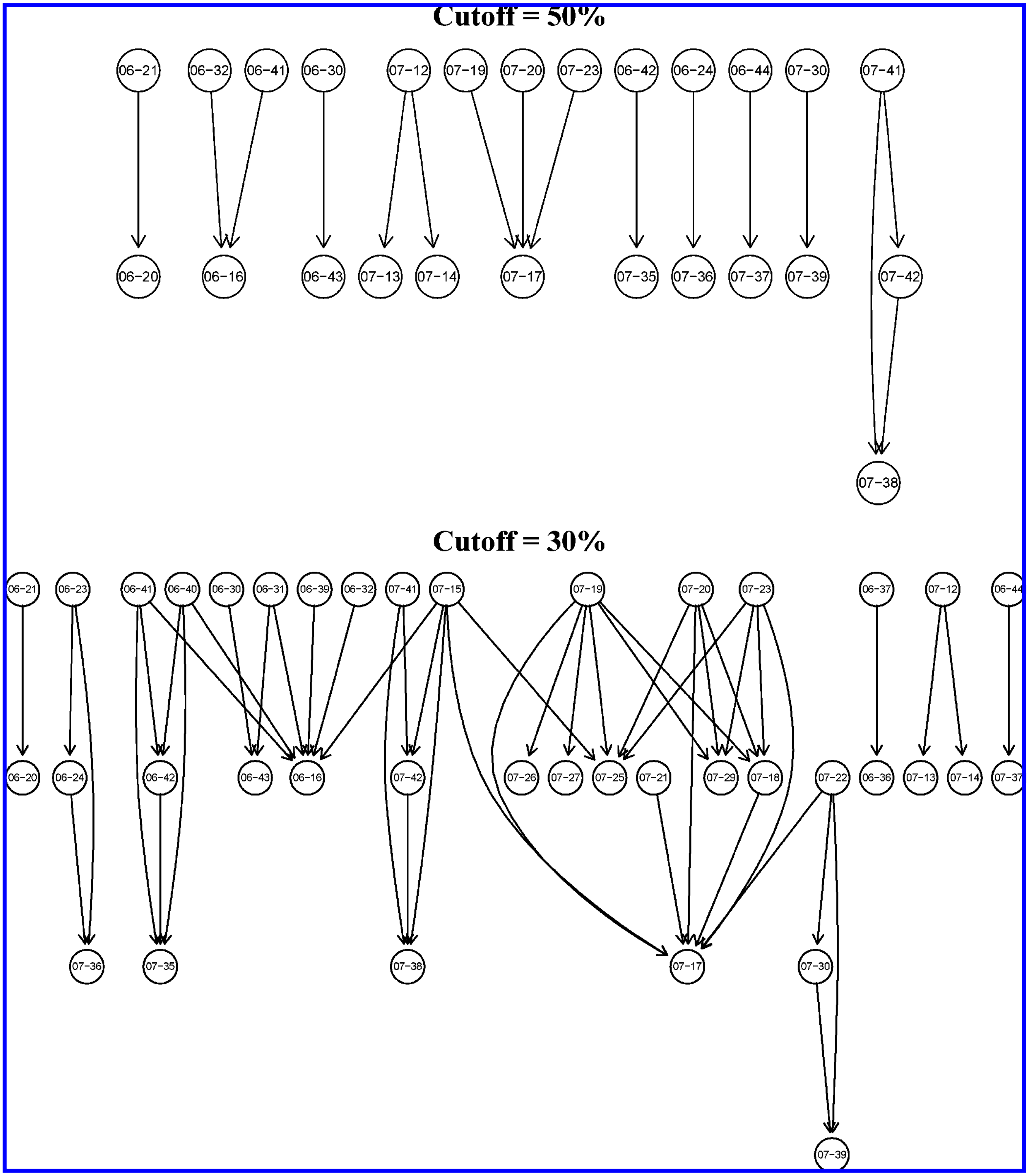
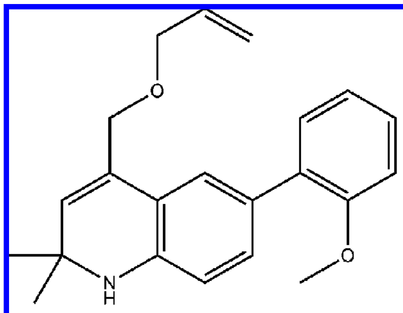


Figure 4. SALI networks for the glucocorticoid inhibitor data set. The cutoffs are reported as a percentage of the range of SALI values for the data set.

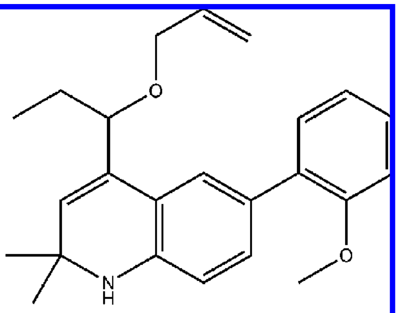
and progesterone receptors. For the purposes of this study we restricted ourselves to the values reported for the glucocorticoid receptors. A number of IC_{50}s were reported as censored values (>2000). In such cases, we entered the censored values as exact values (an egregious assumption inappropriate for more fine-grained modeling but acceptable for a coarse-grained look at the data as provided by the SALI analysis).

Figure 4 shows SALI networks using cutoffs of 50% and 30% of the range of SALI values. The 50% graph appears relatively simple, with a number of disconnected nodes. We initially focus on the compounds **07-17**, **07-20**, and **07-23**. In the 50% graph the latter two compounds are connected to **07-17**. The structures of these compounds are shown in Table 3. It is evident that going from an allyl or phenethyl group at the C4 α-position to an ethyl group results in a

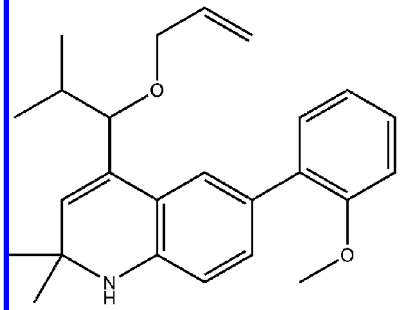
Table 3. Glucocorticoid Inhibitors[a]



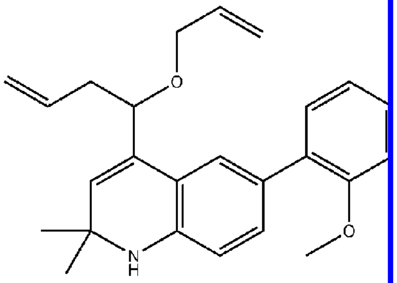
07-15 (2000 nM)



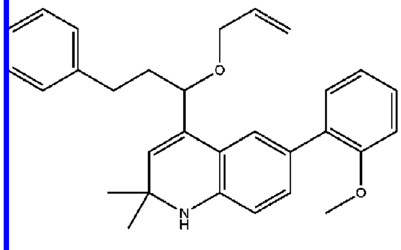
07-17 (355 nM)



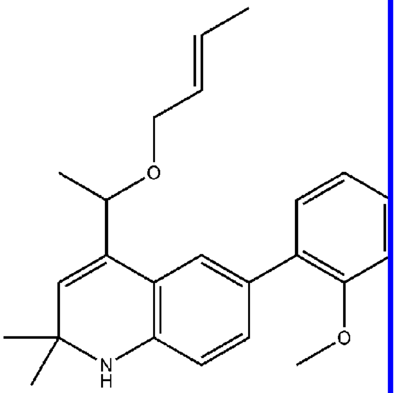
07-18 (710 nM)



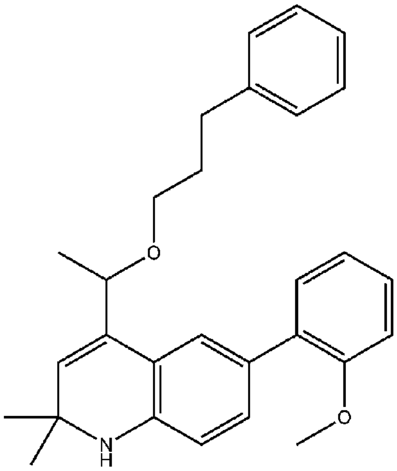
07-20 (2000 nM)



07-23 (2000 nM)



07-25 (210 nM)



07-29 (260 nM)

[a] The values in parentheses indicate the experimentally observed GR IC_{50}.

6-fold increase in activity. This is an example of a broad SAR, as bulky substitutions at this position appear to correlate with activity. Yet, the allylic group is not significantly larger than the ethyl. We are able to gain more insight into the nature of the SAR if we now consider the 30% graph. The subgraph under consideration is now more complex. For example, rather than the direct path **07-20** → **07-17** we now have a path **07-20** → **07-18** → **07-17**. The longer path suggests that not only is the bulk of the substituent an important factor but also the electron density. That is, an increase of pi-electron density in the side chain correlates with a decrease in activity. This obervation is further confirmed if we consider the path **07-23** → **07-18**, where replacement of the phenethyl group by the isopropyl group results in a significant increase in activity. The 30% graph also exhibits another interesting path, viz., **07-15** → **07-17**. From Table 3 we see that rather than decreasing bulk, changing a hydrogen to an ethyl group results in a 6-fold increase in activity. These observations match those made by the authors.[15] However, analysis of the SALI graphs stresses the nonlinear nature of the SAR. That is, simply reducing the bulk at the C4 α-position, which in the limiting case would be a single hydrogen, does not lead to a continuous increase in activity.

The 30% graph highlights a few other SARs. Consider the paths **07-20** → **07-25** and **07-23** → **07-25**. In both cases, we see from Table 3 that by extending the chain of the epoxide, the activity is significantly increased. This is further confirmed by other paths such as **07-20** → **07-29**, in which the allylic group is replaced by a phenethyl group.

Another interesting observation of the SALI graphs considered for this data set is that as we go to more detailed graphs, longer paths that emerge highlight the bulk of the side chains. Though there are many paths which consider changes to the epoxide substitution, these are generally of length 1. In other words, the SALI graphs highlight the fact that changes in the bulk of the side chain at the C4 α-position lead to greater variation in the observed activity. This is also noted by the authors who noted that *"α-substitution has a more significant impact on GR binding affinity than oxygen substitution"*.[15]

As with the melanocortin-4 data set, we also considered pairs of molecules that exhibit significant structural differences yet have very similar activity. An example of such a pair is shown in Table 4. There are two primary differences between the two compounds—a phenylethyl ether moiety on the dihydroquinonline core in **07-33** compared to a thiovinyl moiety in **07-35**. Due to the low similarity they do not appear as connected nodes in the 50% graph. From the above discussion, it is clear that the α-methylation in **07-33** is responsible for its slightly higher activity. Takahashi et al.[15] note that the flourine substitution on C5 in **07-35** is also responsible for a significant increase in potency. In fact if we then compare these two compounds with **07-40,** we see that a combination of α-methylation and flourine substitution leads to a slight decrease in activity, suggesting that the former is key to the potency of these compounds.

4. COMPUTATIONAL CONTROLS

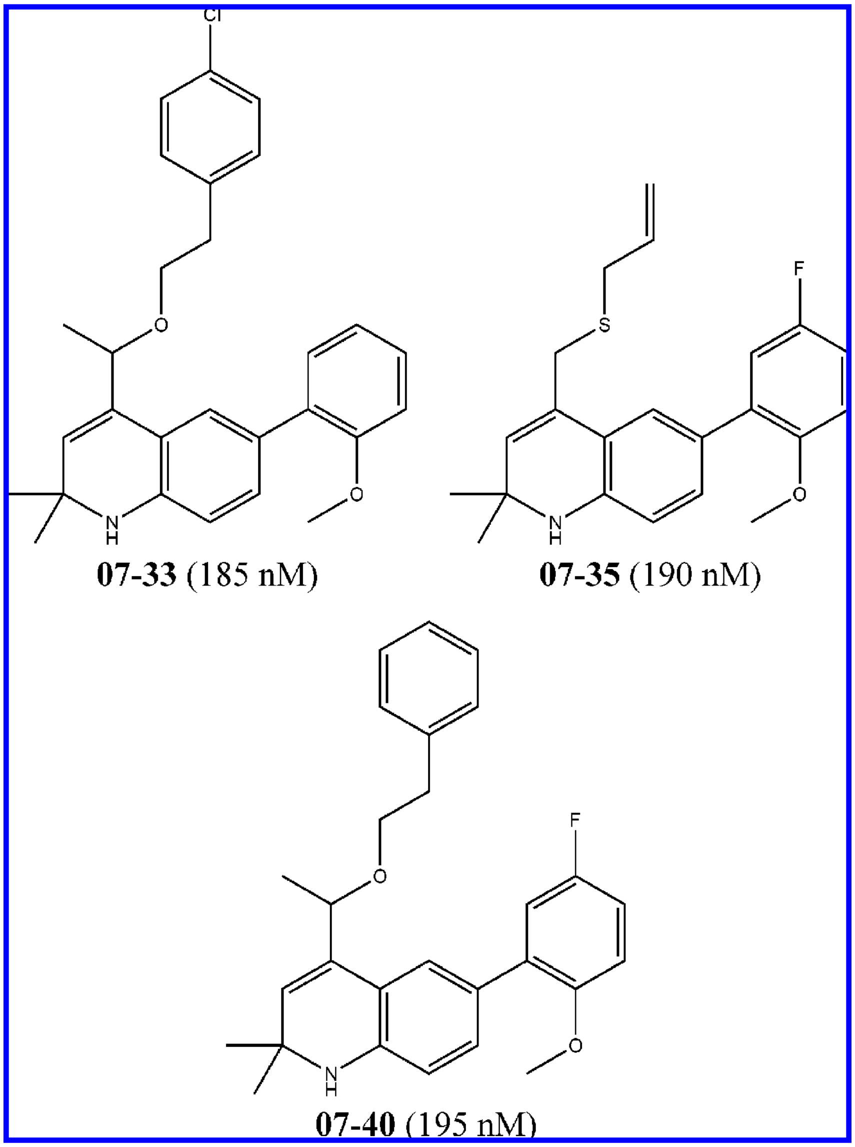
The preceding discussion highlights the utility of the SALI values both for characterizing SARs present in a data set as well as a means to measure the quality of an SAR model. In the following sections we perform a series of computational controls to investigate the behavior of the SALI values in different scenarios, including different similarity metrics, different fingerprint types, and the effects of noisy data.

4.1. Effect of Similarity Metric. To investigate the effect of the similarity metric we consider the PDGFR data set and the BCI fingerprints. We evaluated the SALI matrix using the Tanimoto, cosine, and the Euclidean metrics. It is clear that, numerically, the SALI matrices derived from different similarity metrics will not be identical. However, given that all the metrics can be expected to lead to matrices where the elements have the same ordering, the relative values should be the same. Figure 5 shows a heatmap of the SALI matrix for the three metrics employed, and, as expected, visually there is no difference. We next considered SALI graphs derived from the different SALI matrices. It is obvious then that the SALI graphs, for a given cutoff, will also be identical for different similarity metrics. The underlying reason is that all three metrics considered increase monotonically with increasing similarity. As a result, whatever metric is employed, the relative ordering of the node pairs in the SALI graph will be similar.

4.2. Effect of Fingerprint Type. We next considered how different fingerprint types would affect the SALI values and the resultant SALI graphs. We considered the PDGFR data set and evaluated BCI 1052 bit fingerprints, MACCS 166 bit fingerprints, and the CDK[17] hashed fingerprints of length 1024 bits and 512 bits. Figure 6 displays heatmaps of the SALI matrices generated using BCI, MACCS, and the CDK 1024 bit fingerprints. We did not include the results for the case of the CDK 512 bit fingerprints, since they were identical to the 1024 bit case. It is clear that in the case of the smaller MACCS fingerprints, there is a larger proportion of lighter blocks, indicating the presence of a larger number of "significant" activity cliffs. The heatmap based on the CDK fingerprints is not very different from the BCI-derived plot, though it too has a larger proportion of activity cliffs.

Though it is possible to discern the differences between the three heatmaps, the meaning of the differences is not apparent. Thus we next considered the SALI graphs generated from the three types of fingerprints. Figure 7A−C shows the graphs obtained from each fingerprint type at three different cutoff values. If we compare the graphs for the BCI and MACCS fingerprints, we see that for a given cutoff value the MACCS derived graphs are more complex, with more nodes and edges. This implies that for a given cutoff there are a larger number of pairs of highly similar molecules when using the MACCS fingerprints as opposed to the BCI fingerprints. This is confirmed if we consider the histogram of the similarity values shown in Figure 8. It is clear that the peak of the distribution is shifted toward higher similarity values in the MACCS case. In general, a smaller structural key will exhibit a larger number of "highly" similar pairs than a longer structural fingerprint due to the fact that the small fingerprint encodes more general structural features, whereas the longer fingerprint will tend to encode more specific features.[18,19] As a result, one can expect that for a given molecule there will be more bits set on in the smaller fingerprint than on the larger fingerprint. Since the Tanimoto similarity is essentially a measure of the number of common bits, this implies that the smaller fingerprint will lead to a

Table 4. Glucocorticoid Receptor Inhibitors That Exhibit Similar Activities, yet Are Structurally Similar[a]



07-33 (185 nM) **07-35** (190 nM)

07-40 (195 nM)

[a] The values in parentheses indicate the experimentally determined IC_{50} for the compound.

higher proportion of "high" similar molecules than in the case of the longer fingerprint.

5. DISCUSSION AND FUTURE WORK

One important aspect of generating the SALI networks is that "outlying" compounds (i.e., compounds whose activity is significantly higher than for the rest of the data set) can lead to a network that does not provide much detail. This is because the network is created by considering those compound pairs whose SALI value is greater than a cutoff, where the cutoff is specified as a percentage of the range of the SALI values. The presence of an outlier would cause the cutoff to be skewed toward an artificially high value and so reduce the number of pairs satisfying the cutoff value. One

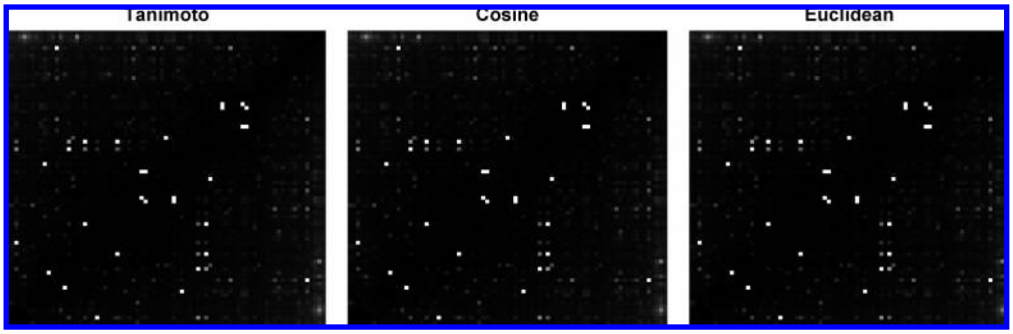


Figure 5. A comparison of the SALI values for the PDGFR data set using the 1052 bit BCI fingerprints and three different similarity metrics.

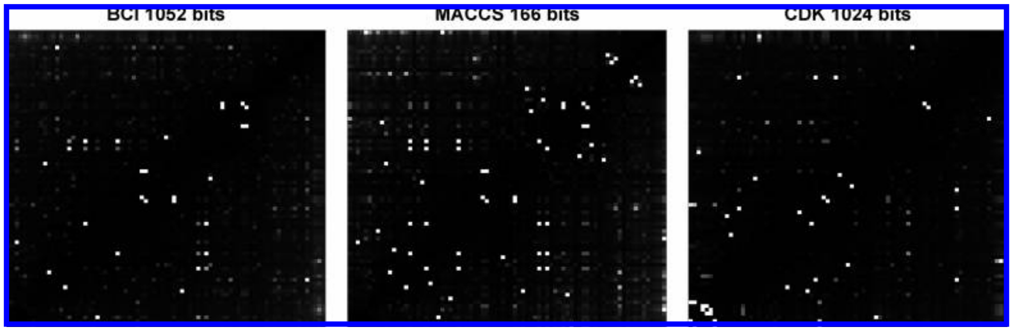


Figure 6. A comparison of the SALI values obtained for the PDGFR data set using three different fingerprint.

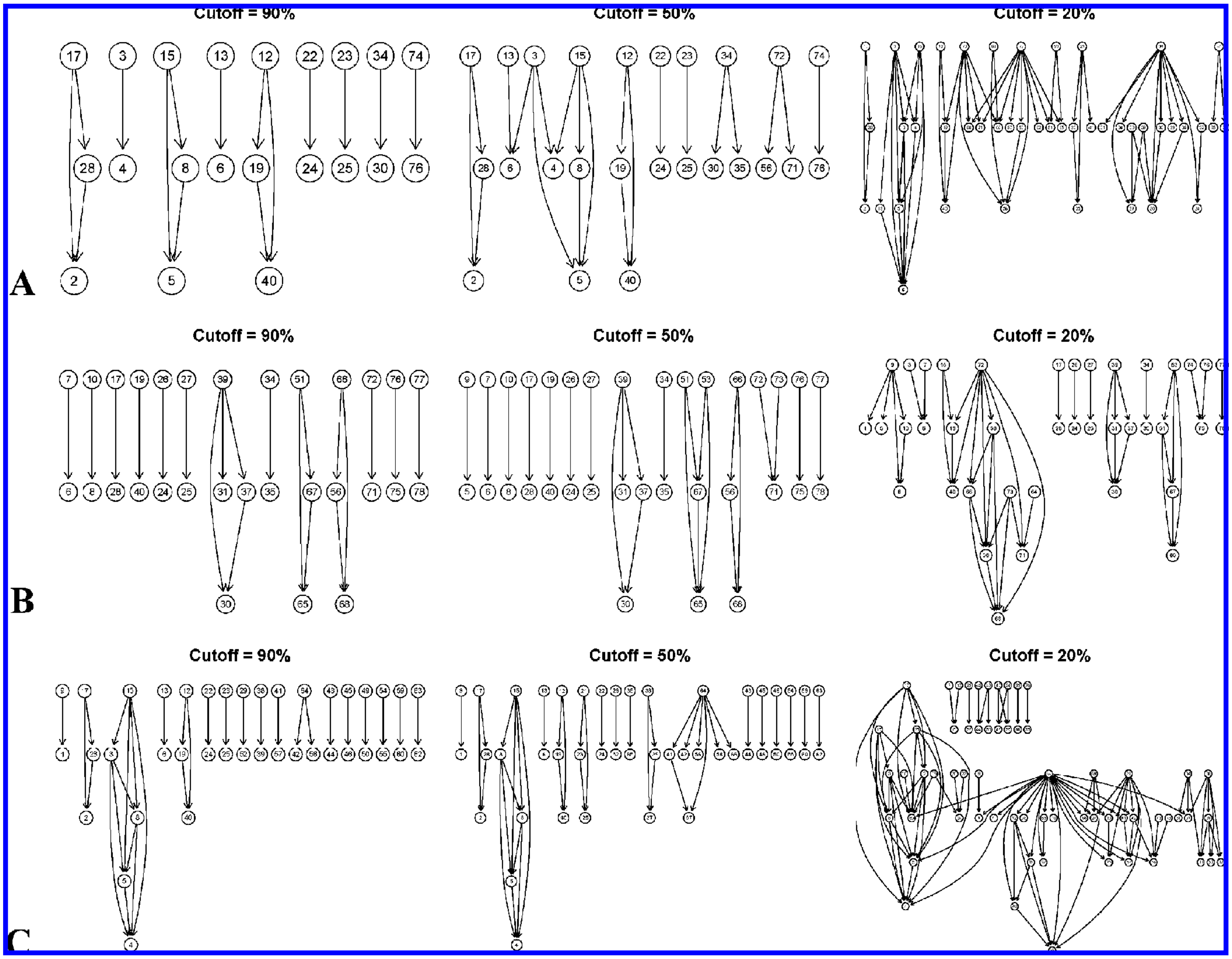


Figure 7. SALI graphs generated for the PDGFR data set using three different fingerprints. **A** represents the graphs generated from BCI 1052 bit fingerprints, **B** from CDK 1024 hashed fingerprints, and **C** from MACCS 166 bit fingerprints. For each fingerprint three different cutoffs were chosen (90%, 50%, and 20% of the range of SALI values).

could modify the SALI metric to take into account outliers by using a quartile based approach. However, we feel that this would lead to a loss of resolution in the final network. Instead, we suggest that a histogram of the activity values be used in combination with the SALI network, so that compounds toward the tails of the distribution can be considered for removal. In an interactive application one would be able to view the network with certain compounds removed from the data set. We note that, even with the current definition of the SALI metric, one can include such outlying compounds in the network yet retain a detailed network structure, simply by using a lower cutoff. The downside is that at such lower cutoffs, the network structure becomes very complex and is difficult to navigate using a static figure.

Related to the problem of outlying compounds is the fact that our approach assumes that activity cliffs identified by the SALI values represent actual structure−activity cliffs. Thus care must be taken to ensure that the experimental values are not in error, since this can lead to a substantially different network. The histogram approach noted can may provide some indication of problematic compounds but obviously does not provide a definitive answer as to whether an outlier is due to experimental error or an actual SAR.

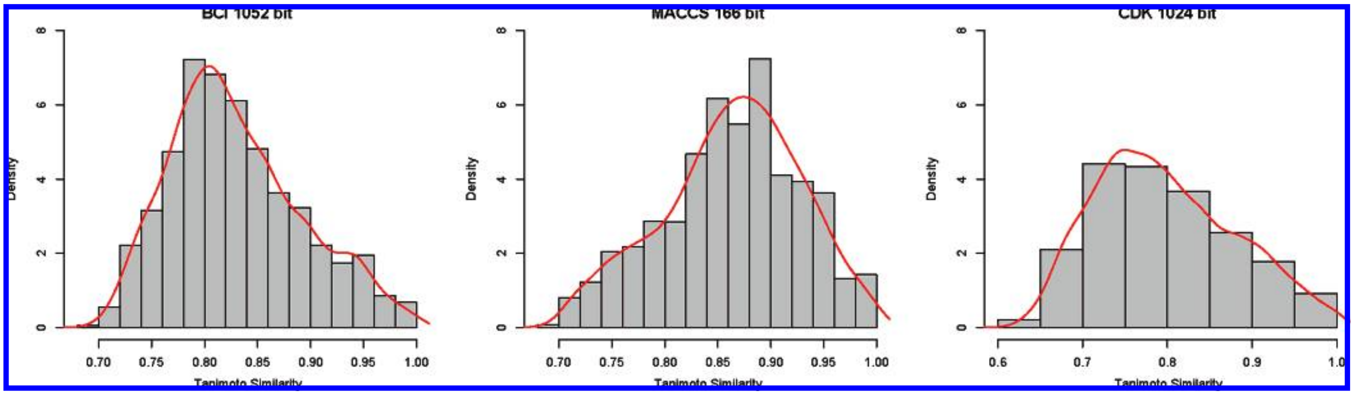


Figure 8. A summary of the distributions of similarities using three different fingerprints with the PDGFR data set.

Our current approach generates graphs using unweighted edges. We have investigated the use of activity differences to weight the edges which does lead to a useful visualization. However there is no restriction on the edge weights. For example, one could also take into account other characteristics such as the synthetic feasibility of going from the less active to the more active member of a pair. If this were numerically quantified, one could then add this as a weight to the edges of a SALI network, allowing multiple levels of information to be encoded.

One of the disadvantages of our approach is that the calculation of the SALI matrix has a $O(N^2)$ time complexity. This suggests difficulties in applying it to large data sets. But we note that the actual calculation time is dominated by fingerprint generation. Once a set of fingerprints has been generated, the evaluation of the SALI matrix is very fast. For example, the calculation of the SALI matrix (after fingerprints have been generated) takes less than 60 ms for 80 molecules, under 10 s for 4000 molecules, and under a minute for 10 000 molecules using the CDK[17] and Java 1.5, on a MacBook Pro with 1GB RAM. Since it is easy to precalculate and store fingerprints, SALI matrices can be rapidly evaluated for large data sets. Furthermore, once the $N \times N$ SALI matrix has been calculated, one could efficiently store the SALI values themselves in a relational database, allowing one to very rapidly generate SALI networks for different values of the cutoff (naïvely, this is an $O(N)$ operation).

Finally, we note that SALI networks could be used to measure the ability of a QSAR model to encode one or more SAR trends. We consider the term "QSAR model" to broadly include different types of models such as statistical, docking, and pharmacophore models. Returning to Maggiora's provocative title "Why QSAR so often disappoints", we realize that it may be unrealistic to expect models to predict activities highly accurately; it may be more important to a drug designer if a model predicts the cliffs qualitatively, i.e., for any pair of molecules, to predict which is more potent. Since each edge of a network is an ordered (in terms of activity) pair of molecules, the edge information can be used measure the quality of the model. Rather than considering the numerical quality of a model (say root mean squared error or r^2), we propose that a model that is able to correctly predict more edges (i.e., the ordering of compounds in an edge) of a SALI network is "better" than a model that predicts fewer edges correctly. In this context, the term "better" implies that a model has been able to better encode the SARs in a data set. Furthermore, one would expect that a model should correctly predict the most significant cliffs with high accuracy, which would decrease as we consider succesively less significant cliffs. Measuring the ability of a model to correctly predict the edges of a SALI network suggests that this approach can be used to design feature selection routines that will result in models that are optimized both for statistical quality as well as their ability to capture the significant activity cliffs. We will present a more rigorous analysis in a forthcoming publication.

6. CONCLUSIONS

SAR modeling generally assumes that small changes in structure lead to small changes in activity (indeed, this is often codified as the "similarity principle"—similar molecules tend to have similar activities). But in many cases, a small structural change can lead to a significant change in activity. These latter instances are termed "activity cliffs" and are often the most interesting parts of an SAR. We have presented an approach to the numerical quantification of such cliffs in terms of the structure−activity landscape index (SALI). Given a data set, one can evaluate the SALI values for each pair of compounds, resulting in a SALI matrix. The matrix can be visualized in a traditional heatmap form. However, we present a novel visualization, whereby the SALI matrix is converted to a network. This representation utilizes a cutoff value, such that the network highlights significant cliffs. As a result, the network view is able to simplify the SARs present in a data set, allowing one to quickly focus on the most significant ones. Though the current work visualizes the networks using static images, we provide an application that allows generation and interactive exploration of SALI networks, coupled with structure depictions and activity information. Furthermore, since the edges in a network represent orderings of pairs of molecules (in terms of activity) we propose that they can be used to quantify the ability of a QSAR model to encode the SARs in a data set.

ACKNOWLEDGMENT

We would like to thank the anonymous reviewers for their helpful comments and suggestions.

Supporting Information Available: Set of R funcions that allows one to generate a SALI matrix from a similarity matrix and activity data and a static network visualization of the SALI matrix for different cutoff values and instructions on how to open and use the file. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Maggiora, G. M. On Outliers and Activity Cliffs−Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.

(2) The initial idea of the SALI was devised by a group at Upjohn Labs in the mid-1990s. (J. Van Drie, M. Lajiness, M. Johnson, T. Hagadone, and G. Maggiora) but was not published; that joint work was publicly disclosed initially by Van Drie in his talk at an IBC conference in San Diego in 1997. It was implemented in the Upjohn cheminformatics system Cousin, now Pfizer's Chemlink. It was also implemented in the VERDI cheminformatics system at Vertex. An apparently independent rediscovery of this idea occurred to a group at Merck in the U.K., also unpublished, and was implemented in a software system at that now-defunct site. Many institutions have implemented this idea as a protocol for the commercial cheminformatics workflow tool, Pipeline Pilot.

(3) Leach, A.; Jones, H.; Cosgrove, D.; Kenny, P.; Ruston, L.; MacFaul, P.; Wood, J.; Colclough, N.; Law, B. Matched Molecular Pairs as a Guide in the Optimization of Pharmaceutical Properties; a Study of Aqueous Solubility, Plasma Protein Binding and Oral Exposure. *J. Med. Chem.* **2006**, *49*, 6672−6682.

(4) Agrafiotis, D.; Shemanarev, M.; Connolly, P.; Farnum, M.; Lobanov, V. SAR Maps: A New SAR Visualization Technique for Medicinal Chemists. *J. Med. Chem.* **2007**, *50*, 5926−5937.

(5) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571−5578.

(6) Digital Chemistry, last accessed September 2007.

(7) 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.

(8) Gentry, J.; Carey, V.; Gansner, E.; Gentleman, R. Laying Out Pathways With Rgraphviz. *R. News* **2004**, *4*, 14−18.

(9) Ellson, J.; Gansner, E.; Koutsofios, L.; North, S.; Woodhull, G. Graphviz - Open Source Graph Drawing Tools. *Graph Drawing* **2002**, *2265*, 483−484.

(10) Pietriga, E. A Toolkit for Addressing HCI Issues in Visual Language Environments. *IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC)* **2005**, *1*, 145−152.

(11) Pandey, A.; Volkots, D. L.; Seroogy, J. M.; Rose, J. W.; Yu, J.-C.; Lambing, J. L.; Hutchaleelaha, A.; Hollenbach, S. J.; Abe, K.; Giese, N. A.; Scarborough, R. M. Identification of Orally Active, Potent, and Selective 4-Piperazinylquinazolines as Antagonists of the Platelet-Derived Growth Factor Receptor Tyrosine Kinase Family. *J. Med. Chem.* **2002**, *45*, 3772−3793.

(12) Tran, J. A.; Chen, C. W.; Jiang, W.; Tucci, F. C.; Fleck, B. A.; Marinkovic, D.; Arellano, M.; Chen, C. Pyrrolidines as Potent Functional Agonists of the Human Melanocortin-4 Receptor. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 5165−5170.

(13) Tran, J. A. et al. Pyrrolidinones as Orally Bioavailable Antagonists of the Human Melanocortin-4 Receptor with Anti-Cachectic Activity. *Bioorg. Med. Chem. Lett.* **2007**, *15*, 5166−5176.

(14) Takahashi, H.; Bekkali, Y.; Capolino, A. J.; Gilmore, T.; Goldrick, S. E.; Nelson, R. M.; Terenzio, D.; Wang, J.; Zuvela-Jelaska, L.; Proudfoot, J.; Nabozny, G.; Thomson, D. Discovery and SAR Study of Novel Dihydroquinoline Containing Glucocorticoid Receptor Ligands. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 1549−1552.

(15) Takahashi, H.; Bekkali, Y.; Capolino, A. J.; Gilmore, T.; Goldrick, S. E.; Kaplita, P. V.; Liu, L.; Nelson, R. M.; Terenzio, D.; Wang, J.; Zuvela-Jelaska, L.; Proudfoot, J.; Nabozny, G.; Thomson, D. Discovery and SAR Study of Novel Dihydroquinoline Containing Glucocorticoid Receptor Agonists. *Bioorg. Med. Chem. Lett.* **2007**, *17*, 5091−5095.

(16) Chen, W.; Wang, W. The Use of Bioisosteric Groups in Lead Optimization. *Annu. Rep. Med. Chem.* **1986**, *21*, 283−291.

(17) Steinbeck, C.; Hoppe, C.; Kuhn, S.; Floris, M.; Guha, R.; Willighagen, E. Recent Developments of the Chemistry Development Kit (CDK) - An Open-Source Java Library for Chemo- and Bioinformatics. *Curr. Pharm. Des.* **2006**, *12*, 2110−2120.

(18) Swamidass, S. J.; Baldi, P. Mathematical Correction for Fingerprint Similarity Measures to Improve Chemical Retrieval. *J. Chem. Inf. Model.* **2007**, *47*, 952−964.

(19) Guha, R.; Dutta, D.; Jurs, P.; Chen, T. R-NN Curves: An Intuitive Approach to Outlier Detection Using a Distance Based Method. *J. Chem. Inf. Model.* **2006**, *46*, 1713−1722.

CI7004093