Published in final edited form as:

Methods Mol Biol. 2011 ; 672: 101–117. doi:10.1007/978-1-60761-839-3_3.

The Ups and Downs of Structure-Activity Landscapes

Rajarshi Guha

NIH Chemical Genomics Center Rockville, Maryland 20852

Abstract

In this chapter we discuss the landscape view of structure-activity relationships (SARs). The motivation for such a view is that SARs come in a variety of forms, such as those where small changes in structure lead to small changes in activity or where small structural lead to significant changes in activity (also termed activity cliffs). Thus, an SAR dataset is viewed as landscape comprised of smooth plains, rolling hills and jagged gorges. We review the history of this view and early quantitative approaches that attempted to encode the landscape. We then discuss some recent developments that directly characterize structure-activity landscapes, in one case with the goal of highlighting activity cliffs while the other allows one to resolve different *types* of SAR that may be present in a dataset. We highlight some applications of these approaches, such as predictive model development and SAR elucidation, to SAR datasets obtained from the literature. Finally, we conclude with a summary of the landscape approach and why it provides an intuitive and rigorous alternative to standard views of structure-activity data.

Keywords

QSAR; glucocorticoid; melanocortin; activity cliff

1 Introduction

The analysis of Structure-Activity Relationships (SAR) is one of the fundamental tasks in medicinal chemistry. A variety of methods have been described for such analyses, ranging from simple visual inspection of compounds and their activities, scaffold analysis to more quantitative methods such as Quantitative Structure-Activity Relationship (QSAR) models.

The fundamental idea underlying all these approaches is that one somehow systematically identifies structural differences between molecules that lead to differences in their activities. In QSAR methods, this is achieved by correlating numerical descriptions of structural features to the activities of the molecules. A simpler, more visual procedure is to analyse the molecules in a pairwise fashion and highlight structural differences between pairs of molecules and the corresponding differences in their activities. The simplest form is to generate R-group tables, though this can become cumbersome when there are multiple R-groups present in a series of molecules. While these approaches are useful for small collections, one would like to extend this approach to a more automated system, that could handle much larger collections, such as those from high throughput screens. Thus various computational schemes have been developed that try to capture the structure-activity information encoded in pairs of molecules. An early approach is the use of “neighborhood behavior” described by Patterson et al[1], which simply plotted the differences in biological activity versus the differences in some descriptor for pairs of molecules. While useful, the method does not really provide insight into SAR trends, which is not surprising since it was designed for diversity analysis of libraries. A more recent example of pairwise analyses is the idea of matched molecular pairs [2]. In that work, Leach et al note that traditionally, identifying pairs of molecules to highlight how certain structural changes lead to

improvements (or degradation) in activity has usually been a manual and subjective process. They described a statistical method that provides a probability that indicates that a certain structural change will cause a change in some property (such as aqueous solubility) in a specific direction. Another approach, similar in intent, focuses on identifying molecular transformations to characterize local QSAR's [3]. This approach consists of two methods, one for analyzing molecular pairs and deriving the structural transformations that capture the differences in their activities and another method that can transform a query molecule into other related structures based on transformations observed in a previously studied collection.

In all the approaches noted above there is an implicit assumption that as one moves from one molecule to another, via a structural change, one observes a change in activity. This suggests that one could view a collection of molecules and their activities as a “landscape”. For simplicity, if we consider a 3D landscape, then the X and Y axes would be defined by some structural descriptor (or combinations of more than 2 descriptors) and the Z axis would represent the activity. An example is shown in Figure 1. The surface is generated by evaluating a set of molecular descriptors for a collection of 81 inhibitors [4] of the melanocortin-4 receptor and then performing a principal components analysis. The first two principal components were plotted on the X and Y axes, and a scaled form of their IC_{50} 's on the vertical Z axis (activity increasing in the upwards direction). Note that the actual surface is a quadratic interpolation of the 81 points.

As expected the surface is not flat. More importantly, the peaks and troughs of the surface highlight interesting aspects of the SAR. More specifically, if one considers the peak, it represents a molecule with very good activity. But, just in front, we see a relatively low portion of the surface corresponding to molecules with poorer activity. The key thing to note is that these two sets of molecules are close to each other on X and Y axes, i.e., the descriptor space. Hence they are structurally similar, but show a significant difference in activity. Such cases, have been termed “activity cliffs” by Maggiora [5] and could be considered as the most interesting parts of an SAR. This view also highlights why QSAR models, especially those based on machine learning do not always perform very well. These models attempt to encode the landscape. However, the presence of activity cliffs is problematic since they represent discontinuities - which by definition cannot be encoded reliably by a machine learning model.

The landscape view of the SAR's for a collection of molecules provides an intuitive way to visualize the relationship between the changes in their structures and the resulting differences in their activities. However, it is also useful to be able to quantify the landscape. In other words, given the landscape view, can we develop metrics that will allow us to quantify the roughness (or smoothness) of the surface. Note that the SAR landscape will, in general, be a high-dimensional hypersurface. As a result, while the 3D visualization of such surfaces is intuitive, methods to characterize such representations should take the high dimensionality into account.

1.1 Outline

The remainder of this chapter discusses recent developments in the characterization of structure-activity landscapes. Section 2 presents the formal definitions for two recently described functions aimed at characterizing SARs and activity cliffs. Section 3 describes some applications and features of these approaches. Finally, Section 4 summarizes the current state of the art.

2 Definitions

SAR landscapes can be characterized in a variety of quantitative ways. Which way is preferred is problem dependent. Thus, if one is interested in directly identifying activity cliffs, one could employ the Structure-Activity Landscape Index (SALI). On the other hand, if one is more interested in characterizing SAR trends, the Structure-Activity Relationship Index (SARI) may be preferable. While the focus of this chapter is the description of SALI, we briefly cover the definitions of both these indices.

2.1 Structure-Activity Landscape Index

The SALI [6] is defined as

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

where A_i and A_j are the activities of molecules i and j and $\text{sim}(i, j)$ is the structural similarity between the two molecules. While any form of structural similarity measure can be used for the denominator, most work has employed a Tanimoto similarity coefficient between two binary fingerprints. Of course, for identical molecules (and also those pairs of molecules whose fingerprint representations are identical), the above measure is undefined. For a collection of molecules one can evaluate a symmetric “SALI matrix”, in which elements with very large values correspond to activity cliffs. A useful visual summary of the SALI values is to simply plot the matrix as an image, as shown in Figure 2. In this plot, the axes correspond to molecules and have been ordered such that the more active molecules are located to the right and top. As a result, white blocks immediately identify pairs of molecules making up activity cliffs. One can further prioritize such cases, by noting that white elements in the top right correspond to an active molecule which by a small structural change becomes another active molecule. While it is nice to see that the change does not decrease activity, it is not as beneficial as having an *inactive* molecule become an active molecule by a small structural change, as represented by white elements in the bottom right.

While the matrix visualization is a useful summary, it is rather limited and static in nature. We have previously described an alternative visualization of the SALI matrix, that lends itself to a more dynamic view of the SALI data, and as will be described in the following sections, presents a powerful tool to characterize SAR datasets and models.

As described in Ref. [6], if we consider each molecule to be a node, we can construct a graph such that an edge is drawn between molecules i and j if $\text{SALI}_{i,j} > C$ where C is some arbitrary cutoff. Clearly, with smaller values of C , more and more molecules will be connected leading to a relatively complex network. However, at high values of C , only those molecular pairs that have very high SALI values will remain in the network. In other words, one can use a high value of C to quickly “zoom in” on the most significant activity cliffs. An example of such a SALI network is shown in Figure 3. The network is laid out such that the edges are directed with the molecule at the tail having poorer activity than the molecule at the head. In addition, as one goes from top to bottom of the graph, the activity improves. It should be noted that even though some nodes are at the same horizontal level, they do not have identical activities. This aspect is an artifact of the layout algorithm.

While the SALI network is a useful way to explore series of structural changes that lead to improvements in activity (especially with the interactive tool described in Ref. [6]), it also allows us to integrate a variety of external information. A simple example is to overlay synthetic feasibility in going from one molecule to another. Thus for every pair of

molecules, one can evaluate a synthetic feasibility score (SFS) that characterizes the ease with which a molecule can be synthesized from another. Since most SFS's consider individual molecules [7, 8], a trivial way to consider pairs of molecules is to simply take the difference of the SFS's of the two molecules in question. This value can then be used to color code the edges of the SALI network.

While the definition above employs IC_{50} 's, one can replace it with any value that tracks activity. Thus, one could employ docking scores in the numerator of Equation 1, the assumption being that docking scores correlate to binding affinity (though this is not always the case). Rather than employing surrogates for activity, one could also consider employing entire dose response curves in place of single point activities, as described in Section 3.1.

2.2 Structure-Activity Relationship Index

An alternative approach and more general approach to characterizing SAR's in a dataset was described by Peltason and Bajorath [9] in which they defined the SAR Index (SARI). The approach is based on the categorization of SARs into continuous, discontinuous and heterogenous, such that the SARI quantifies these categories when applied to a given dataset and target. In their terminology, a discontinuous SAR would correspond to the presence of activity cliffs, whereas a continuous SAR would correspond to smooth regions of the SAR landscape (where a small change in structure leads to little or no change in activity). Mathematically, the SARI is defined as

$$SARI = \frac{1}{2} (\text{score}_{\text{cont}} + 1 - \text{score}_{\text{disc}}) \quad (2)$$

where $\text{score}_{\text{cont}}$ (the continuity score) measures the potency weighted structural diversity (based on pairwise ligand similarity calculations) of a group of compounds and $\text{score}_{\text{disc}}$ (the discontinuity score) measures the average potency difference between pairs of compounds. It is important to note that the SARI value is defined on a *set* of compounds, rather than individual pairs (though the latter are employed in the calculation of the SARI value for a collection). In general, the SARI values will range from 0 to 1, where higher values correspond to more continuous SAR's and smaller values to discontinuous SAR's (i.e., activity cliffs). The detailed mathematical derivations of these scores can be found in the original work. The authors then apply this methodology to a variety of compounds classes targeting a variety of receptors. The results indicate that the SARI values are able to characterize the SAR's present in each of these datasets, highlighting smooth as well as discontinuous portions of the individual SAR landscapes. One of the drawbacks of this approach is that it is based on pre-specified cutoff values. For example, the discontinuity score employs the Tanimoto similarity value between pairs of molecules, but only for those pairs whose T_c crosses a threshold of 0.6. While they report that increasing the threshold value did not significantly change the final results, this is still an arbitrary constant. An example, taken from the original work is shown in Figure 4. The four structures are taken from a collection of thromboxane synthase inhibitors. The numbers on the edges indicate the Tanimoto similarity between the structures. As reported, the SARI value for the set of 23 compounds was 0.46 which is due to a mix of continuous and discontinuous SAR's present in this dataset (highlighted in Figure 4). Due to the presence of both types of SAR's in a single dataset, the authors characterize this as a *heterogenous relaxed* SAR.

3 Case Studies

3.1 Richer activity data

Equation 1 above employs single point activities. While informative, it is a one-dimensional view of the activity of a molecule. Specifically, it describes the potency but does not provide any information regarding the efficacy of the molecule. This information can be obtained from dose-response assays. In such cases, one can obtain a dose-response curve, which is generated by fitting the Hill equation [10].

$$y = S_0 + \frac{S_{inf} - S_0}{1 + 10^{(-\log AC_{50} - x)n}} \quad (3)$$

One now has four parameters describing the biological activity of the compound: S_0 , S_{inf} , h and $\log AC_{50}$. Figure 5 shows a dose-response curve fit using the Hill equation, highlighting the meaning of the four Hill parameters. S_0 and S_{inf} represent the activities at the initial and final portions of the curve. The AC_{50} is the concentration at which one observes 50% activity. Note that the curve fit provides us with $\log AC_{50}$ from which we calculate the AC_{50} . Finally, the slope of the curve between initial and final plateau, is denoted by h .

As a result, rather than simply use the activity in the numerator of Equation 1, one could employ the four Hill parameters. Thus, Equation 1 could be rewritten as

$$SALI_{i,j} = \frac{D(\{P_i\} - \{P_j\})}{1 - sim(i, j)} \quad (4)$$

where $D()$ represents the Euclidean distance function and $\{P_i\}$ and $\{P_j\}$ are the Hill curve parameters for molecules i and j .

When the SALI matrix is regenerated using Equation 4, we get very similar results to that obtained using Equation 1. While the most significant cliffs (i.e., largest values in the SALI matrix) are identical in both cases, the use of curve parameters leads to some differences in the less significant cliffs. Compared to SALI matrix obtained using IC_{50} values only, a number of compound pairs are identified as cliffs. Figure 6 shows the SALI matrices derived from these two approaches for a set of 96 compounds tested in an ERK phosphorylation assay.

3.2 SAR landscapes and predictive models

While the use of SALI networks to explore landscapes is useful, one can employ them for other purposes. Guha and Van Drie [11] described the use of SALI networks as a means of measuring the quality of Quantitative Structure-Activity Relationship (QSAR) models.

The key to this application is to realize that the SALI network is directed. Each pair of connected nodes is thus ordered. Given a predictive model, and a SALI network at a given cutoff, one can use the predicted values from the model and identify how many of the edges in the network are correctly ordered. As noted above, for large values of the cutoff the network highlights the most significant activity cliffs. Thus one can characterize how well a model predicts the most significant activity cliffs. But this is only a partial view of the predictive ability of the model. Instead, we perform this operation on SALI networks generated at a series of cutoffs. The original work evaluated cutoffs from 0% to 100% of the maximum SALI value for the dataset. The result of this is that one can plot the number of correctly predicted edges versus the cutoff. An example of such a plot is shown in Figure

7A, where X represents a fraction of the maximum SALI value for the dataset, and $S(X)$ represents a normalized version of the edge count, such that $S(X) = 0$ indicates that half the edges are predicted correctly, $S(X) = 1$ indicates all the edges are predicted correctly and $S(X) = -1$ indicates that all the edges are mispredicted. The curve has two important parts - the region around $X = 0$ and the region around $X = 1$. At $X = 0$, all pairs of molecules are connected and thus the number of edges correctly predicted at this point, is essentially an ordering of the entire dataset. That is, the value of $S(0)$ represents the ability of the model to capture the landscape as a whole. In this sense, $S(0)$ is conceptually equivalent to the root mean square error (RMSE) of the model, and indeed, in simulations, the correlation between these two values is greater than 0.9. The value of $S(1)$ on the other hand represents the ability of the model to capture the most significant cliffs. In general, while we expect that a good model will capture a large portion of the landscape, we do not expect that machine learning models will have very high values of $S(1)$, since the most significant cliffs represent discontinuities. On the other hand, alternative forms of QSAR models such as docking or pharmacophore methods should be able to accurately identify such significant cliffs, and indeed, SALI curves for such models have been shown to exhibit relatively high values of $S(1)$ [11].

While a useful visual depiction of a ability of the model to encode an SAR landscape, it can be tedious to inspect curves for many molecules. A simple numerical characterization is to represent a SALI curve by the area enclosed between the curve and the $S = 0$ line, as shown in Figure 7B. This metric is termed the SALI Curve Integral (SCI). This approach allows one to summarize the quality (in terms of the ability to encode the landscape) for multiple models in a similar manner. More importantly, the use of the SALI framework to characterize model quality allows one to compare disparate models, on an equal footing. While one could simply use a correlation coefficient to compare the behavior of a docking model and a linear regression model, this is a relatively simplistic measure of quality (and not very robust). On the other hand, the use of the SCI, while still a single number, implicitly contains more information on how well the model is able to characterize the landscape.

3.3 SALI based model selection

While the use of the SCI is an efficient way to explicitly characterize a models' ability to encode the landscape, it also suggests itself as a way to *choose* models that will better encode structure-activity landscapes. In other words, can the SCI (or some other numerical characterization of the SALI curve) be used as a metric to select a set of features that will lead to a model that can capture the details of a SAR landscape with high fidelity. This is essentially the feature selection problem [12]. A multitude of approaches have been applied to feature selection in QSAR modeling ranging from traditional stepwise approaches to methods such as genetic algorithms, simulated annealing and particle swarm methods. From the point of view of SAR landscapes, it does not really matter which specific method is used. Rather, for any given method, we use the SALI curve as the objective function that is being optimized.

As an example we considered the 81 molecule melanocortin-4 inhibitor dataset [4, 13]. Using a set of topological descriptors, we performed feature selection, using a genetic algorithm, to identify a set of good linear regression models. Initially, we considered an objective function that simply minimized the root mean square error (RMSE). Then we replaced the objective function with one based on the SALI curve. Specifically, the we defined the function as

$$S(0) + \frac{1}{2}(S_{\max} - S_{\min}) \quad (5)$$

where $S(0)$ is the value of the SALI curve at $x = 0$ (i.e., the portion of the curve that is analogous to the RMSE for the whole dataset), $S_{\max}$ and $S_{\min}$ are the maximum and minimum values of the SALI curve. This approach was chosen based on the fact that while maximizing the plateau region of the SALI curve would lead to models capturing more of the significant activity cliffs, it would also likely lead to overfitting - since the most significant cliffs would correspond to discontinuities which would have to be “memorized” by the model. Using these two objective functions we then identified a set of 3- and 6-descriptor models. Figure 8 displays a summary of the RMSE's of the models obtained from the two approaches. While it is clear that the SALI based approaches lead to slightly poorer overall RMSEs, the resultant models do have better SCI values compared to the models obtained purely based on optimizing the RMSE. Thus for the 3-descriptor models obtained by optimizing RMSE, the SCI values range from 0.004 to 0.17 whereas for the SALI derived models, the SCI values range from 0.31 to 0.46. While not very good values, this is not surprising as the models themselves do not exhibit significant predictive ability. For the 6-descriptor RMSE derived models the SCIs range from 0.0 to 0.16 whereas for the SALI derived models we observe SCIs from 0.41 to 0.58. Clearly, feature selection based on the SALI curves does lead to models with better ability to encode the SAR landscape. But this comes at a cost. By requiring that models encode at least some of the significant activity cliffs, this introduces some degree of overfitting. This downside could trivially be balanced by modifying the objective function to take into account possible overfitting (such as introducing a penalty for poor cross-validation statistics).

3.4 Network analysis of SAR landscapes

As described previously, both SALI and SARI focus on a network representation of a SAR dataset. In the case of SALI, the network view can be thought of as a transformation of the data that allows subsequent analyses. In other words, it does not necessarily encode physical features of the SAR(s) in the dataset. Thus, while one could evaluate a variety of network metrics such as vertex degree distribution, centrality and so on, such values likely do not have physical relevance. Even so, it is interesting to investigate whether a numerical analysis of the structure of a SALI network can allow us to indicate that one or more SARs are present or not. As an example, we consider a SALI network constructed from a set of 62 glucocorticoid [14, 15] inhibitors. The original network is shown in Figure 9 at a cutoff of 30% of the maximum SALI value in the dataset. Clearly, the actual SALI network has a number of disconnected components. Furthermore, the graph appears to exhibit a non-random structure. In other words, the connectivity in the graph appears to be different from what would be observed in a random graph (i.e., the same number of nodes but connected randomly). To investigate this observation we considered random graphs with the same number of nodes as in the original SALI graph, and edges generated according to a Erdős-Rényi model [16]. We then evaluated various metrics such as the degree distribution, closeness [17] and transitivity [18] values. Figure 10 summarizes the degree distributions of the original SALI graph and the 100 random graphs. In the latter case the degree distribution is obtained by taking the mean degree of each node across the 100 random graphs. Clearly, the random graphs show a much lower degree of variation in vertex degree compared to the original SALI network. Similarly, one can compare the closeness values. As with the vertex degrees, this is also a distribution and Figure 11 compares the closeness for the original SALI graph and the same 100 random graphs. It is clear that the network metrics calculated for the SALI graphs are distinct from those derived from random graphs. This observation suggests that one could employ random graphs to determine a p -value for a given SALI

graph, using it as an indication that the structure of a given SALI graph is not due to chance effects.

The SARI approach has been employed to annotate and analyse network-like similarity graphs [19], where the nodes are chemical structures and nodes are connected by an edge if their Tanimoto similarity is greater than some cutoff. In the study by Wawer et al [19], the SARI values were used to scale the size of nodes to highlight the discontinuity of the SAR in those regions of the graph. These SARI encoded network-like similarity graphs have been used to characterize different aspects of a chemical dataset, ranging from pairwise similarity relationships (via the edges) to SAR discontinuity and activity cliffs (via the SARI based scaling of node sizes). While the authors described a variety of control experiments as well as an analysis of the topology of these network-like similarity graphs in terms of clustering, it is clear that these graphs also exhibit non-random structures. Thus one might expect that a similar approach to that described above for SALI graphs, could be profitably applied to these networks to derive a measure of significance.

4 Conclusions

While the concepts underlying SAR landscapes have been employed by practicing medicinal chemists in a multitude of real world projects, the ability to quantitatively characterize these principles opens up a number of useful possibilities.

From the standpoint of exploring SAR trends in collections of molecules, the network approach has been shown to easily identify specific instances of SAR series that consistently improve activity[20]. More importantly, the approach allows rapid visual identification of such series amongst large collections such as high throughput screening (HTS) datasets, where the majority of compounds will have little or no activity. The network paradigm is also a useful approach to obtain rapid visual summaries of synthetic campaigns, allowing users to highlight changes which lead to improvements in activity [6]. In this sense, the SAR landscape view allows one to easily recapitulate useful synthetic changes, effectively guiding future work on those series.

However, the SAR landscape view is not restricted to visual summaries of SAR datasets. As has been described the SALI network view can be applied to quantitative models that attempt to encode the SAR landscape. Specifically, one can view approaches such as SALI curves and the SCI as alternatives to RMSE and R^2 , that allow a more direct characterization of a predictive model (which might be a machine learning model or a physical approach such as docking and pharmacophore models) in terms of their ability to capture details of the landscape. This view also highlights the problems associated with the different types of model approaches. Thus, machine learning based models will, in general, never be able to predict significant activity cliffs with great fidelity, simply because they represent discontinuities in the landscape. From a physical point of view, such discontinuities indicate that some specific structural feature (say a specific hydrogen bonding donor) is responsible for the jump in activity. Unless one happens to identify a molecular descriptor that specifically characterizes this physical feature, traditional machine learning models will not effectively encode the behavior of such molecules. Moreover, even if one were to include such a specific descriptor, it would exist primarily to capture the few cases of significant activity cliffs, and not contribute much to the other patterns present in the dataset. This would explain why traditional feature selection approaches may not identify such descriptors. On the other hand, SALI based feature selection can identify such features, but at the same time introduces a degree of overfitting into the model. But given that activity cliffs are indicative of specific interactions between ligand and receptor, approaches such as docking and pharmacophore models are inherently better at capturing the details of the

landscape. As shown by Guha and Van Drie [11], this is reflected in the SCI values obtained for these types of models. Even in this case, one could profitably employ SCI values to choose between different scoring functions or even explain why specific scoring functions do better than others for a given target. But more generally, the use of landscapes in characterizing quantitative models allows us to compare models built on disparate methodologies on an equal footing.

In conclusion, while the ideas underlying SAR landscapes are not new, the recent developments in quantitative approaches to characterizing such landscapes and applications of these methods to a diverse set of problems ranging from SAR identification to model quality characterization highlight the utility and flexibility of the landscape paradigm.

References

- [1]. Patterson D, Cramer R, Ferguson A, Clark R, Weinberger L. Neighborhood Behavior: A Useful Concept for Validation of "Molecular Diversity" Descriptors. *J. Med. Chem.* 1996; 39:3049–3059. [PubMed: 8759626]
- [2]. 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. [PubMed: 17154498]
- [3]. Sheridan R, Hunt P, Culberson J. Molecular Transformations as a Way of Finding and Exploiting Consistent Local QSAR. *J. Chem. Inf. Model.* 2006; 46:180–192. [PubMed: 16426054]
- [4]. Tran JA, Chen CW, Jiang W, Tucci FC, Fleck BA, 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. [PubMed: 17629702]
- [5]. Maggiora GM. On Outliers and Activity Cliffs—Why QSAR Often Disappoints. *J. Chem. Inf. Model.* 2006; 46:1535–1535. [PubMed: 16859285]
- [6]. Guha R, Van Drie J. The Structure-Activity Landscape Index: Identifying and Quantifying Activity-Cliffs. *J. Chem. Inf. Model.* 2008; 48:646–658. [PubMed: 18303878]
- [7]. Bertz S. The First General Index of Molecular Complexity. *J. Am. Chem. Soc.* 1981; 103:3599–3601.
- [8]. Allu T, Oprea T. Rapid Evaluation of Synthetic and Molecular Complexity for in Silico Chemistry. *J. Chem. Inf. Model.* 2005; 45:1237–1243. [PubMed: 16180900]
- [9]. Peltason L, Bajorath J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* 2007; 50:5571–5578. [PubMed: 17902636]
- [10]. Lehninger, A. *Lehninger Principles of Biochemistry*. 4th ed.. W.H.Freeman; New York, NY: 2004.
- [11]. Guha R, Van Drie J. Assessing How Well a Modeling Protocol Captures a Structure-Activity Landscape. *J. Chem. Inf. Model.* 2008; 48:1716–1728. [PubMed: 18686944]
- [12]. Duch, W. *Feature Extraction: Foundations and Applications*. Vol. Vol. 207. Springer; Berlin: 2006.
- [13]. Tran JA, 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 AJ, Gilmore T, Goldrick SE, Kaplita PV, Liu L, Nelson RM, 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. [PubMed: 17681466]
- [15]. Takahashi H, Bekkali Y, Capolino AJ, Gilmore T, Goldrick SE, Nelson RM, 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. [PubMed: 16386422]
- [16]. Erdős P, Rényi A. On Random Graphs. *Publicationes Mathematicae*. 1959; 6:290–297.

- [17]. Freeman L. Centrality in Social Networks I: Conceptual Clarification. *Social Networks*. 1979; 1:215–239.
- [18]. Wasserman, S.; Faust, K. *Social Network Analysis: Methods and Applications*. Cambridge University Press; Cambridge, UK: 1994.
- [19]. Wawer M, Peltason L, Weskamp N, Teckentrup A, Bajorath J. Structure-Activity Relationship Anatomy by Network-like Similarity Graphs and Local Structure-Activity Relationship Indices. *J. Med. Chem.* 2008; 51:6075–6084. [PubMed: 18798611]
- [20]. Wawer M, Peltason L, Bajorath J. Elucidation of StructureActivity Relationship Pathways in Biological Screening Data. *J. Chem. Inf. Model.* 2009; 52:1075–1080.

Figure 1.
A schematic representation of a SAR landscape for a set of 81 inhibitors of the melanocortin receptor.

Figure 2. A visual summary of the SALI values for a collection of molecules. Lighter colors represent increasingly significant activity cliffs.

Figure 3.
An example of a SALI network.

Figure 4. A set of molecules highlighting the continuous and discontinuous SAR trends characterized by SARI. The edge labels indicate Tanimoto similarity between the corresponding molecules. Taken from Peltason and Bajorath [9].

Figure 5.

A schematic diagram of a dose response curve fit using the Hill equation. The plot is annotated with three of the four Hill parameters. The remaining parameter is the slope of the curve between the two exponential portions of the curve.

Figure 6. SALI matrices obtained using simple IC_{50} values (left) and dose response curve parameters for a set of 96 compounds tested for activity in an ERK phosphorylation assay.

Figure 7.

A SALI curve and characterizing a models' ability to encode an SAR landscape and the SALI Curve Integral (SCI). In both plots, X is the SALI cutoff as a fraction of the maximum SALI value in the dataset and $S(X)$ represents the normalized edge count at a cutoff X .

Figure 8.
A summary of the RMSEs of linear regression models obtained using traditional RMSE based feature selection and SALI curve based feature selection.

Figure 9.

A SALI graph for a set of 62 inhibitors of the glucocorticoid receptor [14, 15], generated at a SALI cutoff of 30% of the maximum SALI value in the dataset.

Figure 10.

A comparison of vertex degree distributions for the original SALI graph (left) and 100 random graphs (right), with the same number of nodes as in the original SALI graph, but with edges determined by the Erdő-Rényi model

Figure 11.

A comparison of vertex closeness distributions for the original SALI graph (left) and 100 random graphs (right), with the same number of nodes as in the original SALI graph, but with edges determined by the Erdős-Rényi model