ELSEVIER

Navigating structure–activity landscapes

Jürgen Bajorath[1], Lisa Peltason[1], Mathias Wawer[1], Rajarshi Guha[2], Michael S. Lajiness[3] and John H. Van Drie[4]

[1] Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, D-53113 Bonn, Germany
[2] School of Informatics, Indiana University, Bloomington, IN 47406, USA
[3] Lilly Research Laboratories, Indianapolis, IN 46285, USA
[4] Van Drie Research LLC, Andover, MA 01810, USA

The problem of how to explore structure–activity relationships (SARs) systematically is still largely unsolved in medicinal chemistry. Recently, data analysis tools have been introduced to navigate activity landscapes and to assess SARs on a large scale. Initial investigations reveal a surprising heterogeneity among SARs and shed light on the relationship between 'global' and 'local' SAR features. Moreover, insights are provided into the fundamental issue of why modeling tools work well in some cases, but not in others.

Any successful voyage of discovery requires excellent navigation. In the era of Christopher Columbus, marine navigation involved maps of doubtful quality, a compass and the ability to gauge one's direction using the sun and stars, along with 'dead reckoning' – an intuitive mariner's sense of position and direction. It is, therefore, not surprising that the invention of the chronometer for reliably determining one's position on the surface of the globe, over 200 years after Columbus, had a profound positive impact on shipping and navigation [1].

In drug discovery, one sometimes has 'maps', for example, the electron density of a ligand–target complex, and 'compasses', including computational models and empirical rules defining the characteristics of 'drug-like' molecules [2–4]. A discovery 'chronometer' has yet to be invented, however, leaving drug-hunters to heavily rely on 'dead reckoning' to this date.

During its early stages, the iterative process of drug discovery mostly focuses on a central question in medicinal chemistry: 'what molecules should be made next?' Key to answering this fundamental question is the relationship between chemical structure and biological activity; compound synthesis must be guided by an analysis of the underlying structure–activity relationship (SAR). An SAR is generally thought to exist if a series of chemical modifications of a core structure results in a defined increase or decrease in biological activity. Accordingly, the central question in medicinal chemistry can be further broken down into a sequence of subsidiary and more specific questions:

- If there is an SAR, what type of SAR do we have?
- Are there trends in the SAR? Are they gradual, coherent and sensible?
- If we have only one scaffold, does this scaffold have the potential to become a drug candidate?
- If we have multiple scaffolds, which one is best poised for further elaboration into a candidate?
- Can we model the SAR? If we have a model, to what extent can we trust it for guidance?

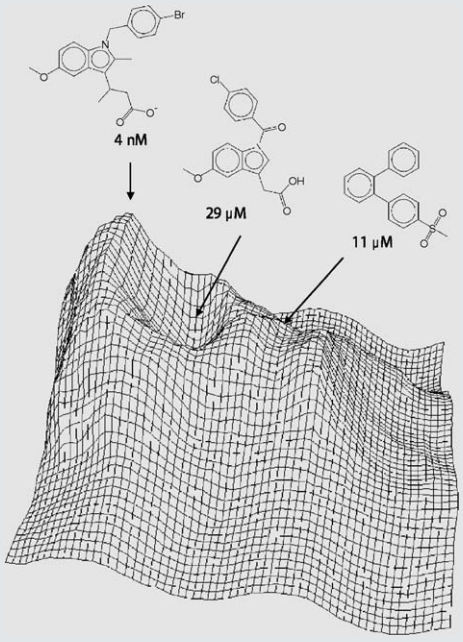
Although it is often difficult for drug discovery teams that need to meet aggressive milestones to pose rigorously and explore such questions, ultimate success in discovery requires doing so. Much like mariners constantly sought to find better methods than dead reckoning, substantial efforts have been made to explore and predict SARs. For example, originating in the 1960s, quantitative SAR (QSAR) modeling has become an integral tool in medicinal chemistry to guide lead optimization efforts [5]; nonetheless, compound selection and optimization are still heavily influenced by intuition and serendipity [6]. It should be noted that chemical intuition is also plagued by its own shortcomings, as was shown in

Corresponding author: Bajorath, J. (bajorath@bit.uni-bonn.de)

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2009.04.003

BOX 1

Activity landscapes and similarity assessment



Cyclooxygenase 2 inhibitors on a schematic activity landscape. The two structures on the left and in the middle represent similar chemotypes but have potency values that differ by several orders of magnitude, thus forming an 'activity cliff'. By contrast, the structure on the right has a distinct chemotype but a potency similar to the second structure, indicating a continuous SAR region.

Variable activity landscapes

Activity landscapes can be rationalized as biological response surfaces in chemical space that are obtained by adding an activity dimension to this space. If we consider a two-dimensional projection of chemical reference space, compound potency is added as a third dimension, thereby obtaining a topological map that is reminiscent of geographical maps. Moving around in the chemical space plane (e.g. making chemical modifications to an active compound) is accompanied by changes in biological activity, giving rise to landscapes with varying topologies.

Representation of molecular structure and assessment of similarity

For computational analysis, chemical structures are usually represented by descriptors encoding molecular structure and properties in a numerical format. The widely used fragment-type descriptors are typically encoded as molecular fingerprints consisting of multiple bit positions reporting the presence or absence of specific structural fragments. The similarity between two molecules is then quantified by comparing their fingerprints A and B using a similarity metric such as the Tanimoto coefficient:

$$Tc(A,B) = \frac{N_{AB}}{N_A + N_B - N_{AB}}$$

Here, N_{AB} is the number of bits set on in both fingerprints and N_A and N_B refer to the number of bits set on in A and B, respectively.

It is important to note that chosen molecular representations and the way compound similarity is assessed influence the topology of activity landscapes and SAR analysis. One of the challenges for the tools and techniques presented herein is to evaluate which of an almost infinite set of possible descriptors and associated similarity metrics is most informative. In addition, it should also be taken into account that, in practice, activity cliffs might turn out to be artifacts resulting from errors in biological assays or testing of samples containing misidentified compounds.

a study on the consistency of medicinal chemists' decision-making [7]. Moreover, the popularity of simple and intuitive 'rules' that suggest how biologically active and drug-like molecules should 'look like' proposes that our understanding of SARs is still rather limited. Consequently, an over-reliance on such empirical rules might lead drug discoverers to disregard promising ideas for new molecules simply because they would not conform to those rules [8].

Activity landscapes

Exploring and understanding SARs are intrinsically difficult because their 'landscapes' can be highly variable (see Box 1). They can be smooth and easily navigated, or jagged and difficult to traverse. One often finds that gently sloped regions where gradual changes in structure are accompanied by moderate changes in activity lead to steep 'activity cliffs' where small modifications in compound structure substantially influence biological effects. It is important to identify and understand activity cliffs because compounds associated with them often indicate chemical features that are crucial for SARs. Such compounds are, however, not only essential to understand SARs, but also they may constitute apparent statistical outliers in QSAR analysis and can unduly influence many calculations in a negative way [9]. Maggiora has suggested [10], provocatively, that one reason why SAR modeling frequently disappoints is that modeling tools do not take into account the possibility that the activity landscape might be different going from one SAR to another. Thus, much like mariners used to 'throw the lead' to look for depth changes in unknown waters, modelers often rely on computational methods that have proven useful to treat an individual SAR. There is no 'SAR weight' that can be thrown, however, and, until recently, no methodology has been available to explore unknown activity landscapes systematically.

SAR analysis functions

To explore SARs on a large scale and better understand how to navigate activity landscapes in a consistent manner, SAR data analysis tools have been developed that systematically relate compound potency and similarity to each other, with an eye toward categorizing different types of SARs. Structure–Activity Similarity (SAS) maps [11] have provided a compelling graphical representation of compound distributions on activity landscapes and a conceptual basis for the development of SAR analysis functions, such as the Structure–Activity Relationship Index (SARI) [12] and the Structure–Activity Landscape Index (SALI) [13]. In Box 2, details of these methods are provided and their application is described in the following.

SALI

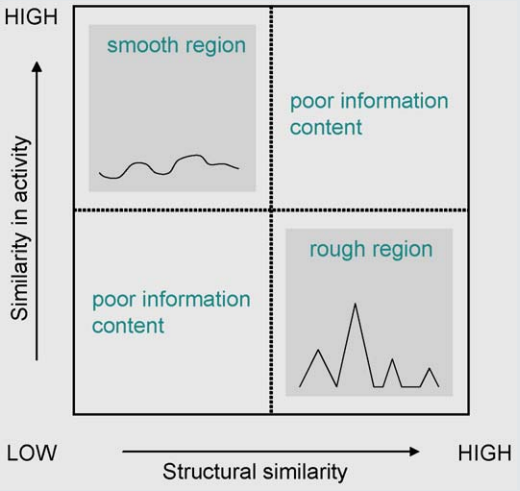
In concert with a medicinal chemist's intuition, a network representation of an activity landscape yields insights into various aspects of the SAR(s). Figure 1 shows two SALI networks derived from published SARs for antagonists of the glucocorticoid receptor

BOX 2

SAR analysis methods

Structure–Activity Similarity (SAS) maps

SAS maps provide a graphical and numerical tool for the analysis of SARs. Structural similarity of active compounds is plotted against an index for activity similarity, revealing different potency–similarity relationships and regions of varying information content. Information-theoretic comparison of different SAS maps provides a quantitative characterization of smooth or rough activity landscapes.



Schematic representation of an SAS map. Based on pairwise comparison of chemical structure and biological activity within a set of compounds, regions of different SAR character can be identified.

Structure–Activity Landscape Index (SALI)

SALI is designed to identify activity cliffs and compounds representing key inflection points on activity landscapes. In a set of compounds, it assigns each compound pair a score that combines their pairwise similarity and the difference between their potency:

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

Here (and in the following) A_i and A_j are the potency values of the ith and jth molecule, respectively, and sim(i,j) is the calculated similarity of the two molecules. Hence, highest SALI values are obtained for compound pairs that have a high degree of structural similarity but large differences in potency and thus correspond to activity cliffs.

Structure–Activity Relationship Index (SARI)

SARI combines two different score components that account for global SAR continuity and discontinuity. The 'continuity score' calculates the potency-weighted mean of pairwise compound dissimilarity and thus emphasizes the presence of structurally diverse and potent compounds.

$$\mathrm{cont}_{\mathrm{raw}} = \underset{\{(i,j)|i\neq j\}}{\text{weighted mean}}\left(\frac{1}{1+\mathrm{sim}(i,j)}\right) = \frac{\sum_{(i\neq j)}(\mathrm{weight}(i,j)1/(1+\mathrm{sim}(i,j)))}{\sum_{i\neq j}\mathrm{weight}(i,j)}$$

$$\mathrm{weight}(i,j) = \frac{A_i A_j}{1+|A_i - A_j|}$$

The 'discontinuity score' is defined as the similarity-scaled average potency difference among ligand pairs that exceed a predefined similarity threshold and have a significant potency difference. Thus, it identifies activity cliffs.

$$\mathrm{disc}_{\mathrm{raw}} = \underset{\{(i,j)|\mathrm{sim}(i,j)|>0.65,|A_i-A_j|>1\}}{\text{mean}}(|A_i - A_j|\mathrm{sim}(i,j))$$

The "raw" score components are normalized, mapped to the value range between 0 and 1, and combined to yield the final SARI.

$$\mathrm{SARI} = \frac{1}{2}(\mathrm{cont}_{\mathrm{norm}} + (1 - \mathrm{disc}_{\mathrm{norm}}))$$

High SARI values indicate continuous, low values discontinuous and intermediate values heterogeneous SARs. The latter SAR category has two subtypes, 'heterogeneous-relaxed' (resulting from high continuity and discontinuity scores) and 'heterogeneous-constrained' (low continuity and discontinuity scores) SARs.

For SAR contributions of individual compounds, a variant of the discontinuity score has been developed that is calculated for a compound and its nearest neighbors:

$$\mathrm{disc}(i)_{\mathrm{raw}} = \underset{\{j\neq i|\mathrm{sim}(i,j)>0.65\}}{\text{mean}}(A_i - A_j|\mathrm{sim}(i,j))$$

The local discontinuity score reflects the potential participation of each compound in local activity cliffs and hence individual contributions to local and global SAR characteristics.

It should be noted that the assessment of molecular similarity, which is a key component of the SALI and SARI functions, always depends, at least to some extent, on the molecular representations and similarity metrics that are utilized. The analyses presented herein are based on molecular fingerprint representations.

[14,15]. In these networks, each node represents a compound. Two compounds are connected by an edge if their SALI value is greater than a user specified cutoff. Edges are directed according to increasing potency. The network in Fig. 1a characterizes the most significant activity cliffs in the dataset. The relatively simple structure of the SALI network highlights the fact that these pairs of compounds represent relatively rare aspects of the SAR. By examining the structures corresponding to the nodes one can begin discerning specific local SARs. Thus, a change from allyl

Examples of SALI networks and SALI curves. The SALI network is derived from the SALI matrix (see Box 2). Nodes represent compounds and two compounds are connected if their SALI value is greater than a cutoff. By using a high cutoff one can generate a network that highlights only the most significant cliff. Progressively lower cutoffs generate more complex networks that include less significant cliffs. **(a)** A SALI network for a set of glucocorticoid antagonists, using a relatively high cutoff to highlight significant cliffs. Two pairs of activity cliffs are highlighted. **(b)** A SALI network for the same dataset using a lower cutoff. **(c)** SALI curves derived for two models based on different datasets. These curves are used to characterize the ability of an SAR model to encode the activity landscape. The quantity $S(X)$ can be derived for any SALI subgraph resulting from a chosen threshold X. It reports the ability of a computational model to predict edges in the SALI graph and is defined as $S(X) = ((n_{\mathrm{correct}} - n_{\mathrm{wrong}})/n_{\mathrm{edge}})$ where n_{correct}, n_{wrong}, and n_{edge} represent the number of correctly predicted edges, incorrectly predicted edges and total number of edges, respectively. A SALI curve is a diagram of $S(X)$ as a function of X (where X ranges from 0 to 1). Ideally, a model that captures the entire landscape will have a curve that rapidly plateaus at 1. The PDGFR curve is derived from a linear regression model and the hERG curve is derived from a CoMFA model. The ability of the 3D method to capture more details of a landscape, compared to a 2D method, is evident and highlights the utility of these curves as a measure of model quality.

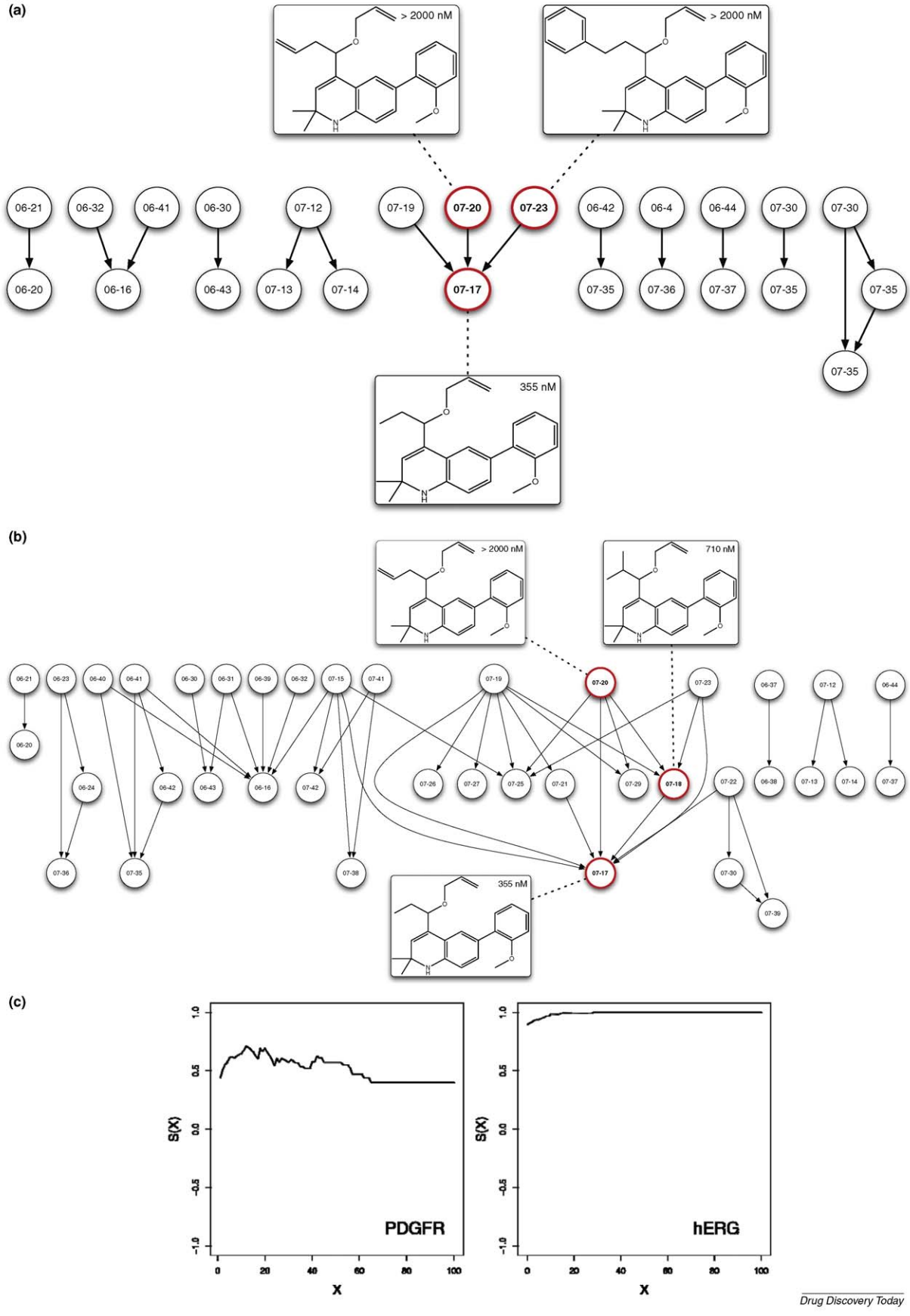

(a)
> 2000 nM
> 2000 nM
06-21
06-32
06-41
06-30
07-12
07-19
07-20
07-23
06-42
06-4
06-44
07-30
07-30
06-20
06-16
06-43
07-13
07-14
07-17
07-35
07-36
07-37
07-35
07-35
07-35
355 nM
(b)
> 2000 nM
710 nM
06-21
06-23
06-40
06-41
06-30
06-31
06-39
06-32
07-15
07-41
07-19
07-20
07-23
06-37
07-12
06-44
06-20
06-24
06-42
06-43
06-16
07-42
07-26
07-27
07-25
07-21
07-29
07-18
07-22
06-38
07-13
07-14
07-37
07-36
07-35
07-38
07-17
07-30
07-39
355 nM
(c)
S(X)
PDGFR
X
S(X)
hERG
X
Drug Discovery Today

FIGURE 1

(07–20, >2000 nM), or phenylethyl (07–23, >2000 nM), to ethyl (07–17, 355 nM), results in a 6-fold improvement in potency (Fig. 1a). One might conclude that reducing bulk at this position might improve activity. Yet, the allyl group is not significantly larger than the ethyl substitution, especially when compared to the bulk of the phenylethyl group. One can then consider a more detailed network, as shown in Fig. 1b. By considering the highlighted nodes, we observe that reduction in pi-electron density of the substituent at the C4-α position improves potency by at least 3-fold from compound 07–20 (>2000 nM) to 07–18 (710 nM) and by 2-fold from compound 07–18 (710 nM) to 07–17 (350 nM). Interestingly, this network indicates that the effect of bulky substituents at the C4-α is not linear. In addition to looking at individual nodes, the analysis indicates that longer paths in the network focus on the bulk of side-chains at the C4-α position.

The SALI map does not infer these trends on its own – it facilitates the process of a chemist identifying them (indeed, the conclusions drawn from Fig. 1 match closely with those made by Takahashi *et al.* in their original studies [14,15]). The strength of this representation is that it highlights key compound pairs, which is especially useful when the SAR is large, for example when studying the results of a high-throughput screen.

When an SAR exists, selected pairs of molecules make it possible to formulate a local hypothesis; for example in the glucocorticoid examples mentioned above, replacement of a bulky substituent with a smaller one significantly improves potency, as does lower pi-electron density in these substituents. Given these tentative hypotheses, one scans the remaining activity cliffs to see if there are other pairs that are consistent with that hypothesis. If so, the hypothesis is further supported. If not, one must re-evaluate the explanation of the original pair or examine the dataset to identify other molecules that could, in principle, test the hypothesis. Of course, this is precisely the process a medicinal chemist follows when intuitively analyzing an SAR. The SALI map, however, augments this natural, intuitive approach in a consistent and unbiased manner.

SALI analysis also helps to assess how systematically and thoroughly an SAR has been explored around a given scaffold and determines whether there are 'holes' in this SAR, that is molecules which have not been made but might be expected to have high potency; or whether there are obvious points of metabolic liability (e.g. benzylic carbons) that might be replaceable while maintaining potency. Furthermore, the presence of sharp activity cliffs in SALI maps suggests that many small variations of a structure should be synthesized to ensure that an activity landscape has been adequately sampled. Also, the presence of nodes with few connections to other nodes in a SALI network might well point to key compound series that have the greatest potential for efficient exploration.

Appropriate SAR modeling

The literature of modeling SARs is vast, ranging from QSAR to pharmacophore and structure-based models. One of the most perplexing things to drug discovery teams relying on such models is that sometimes they are useful, but other times they fail, leaving open the key question 'to what extent can one trust them for guidance?'. SALI networks explicitly address this question in the context of activity landscapes. By studying how well a given model is able to predict the ordering of any pair of molecules in the SALI network a SALI curve is derived [16]. Figure 1c shows a SALI curve for a linear regression model for a set of platelet-derived growth factor receptor (PDGFR) antagonists and for a comparative molecular field analysis CoMFA [17] model for a set of human ether-à-go-go related gene (hERG) inhibitors. The underlying hypothesis is that if such a curve plateaus near values approaching 1, a reliable model might have been applied. This approach is applicable to any type of model (i.e. QSAR, pharmacophore or structure-based). As such, it should be informative in prioritizing alternative models with respect to different SAR characteristics.

SARI

Systematic SARI profiling of many different sets of active compounds has shown that the majority of global SARs are heterogeneous in nature [12], that is their activity landscapes contain both gently sloped regions and activity cliffs. Table 1 reports the SAR categories for a representative ensemble of compound sets active against diverse targets. For medicinal chemistry, the prevalence of globally heterogeneous SARs is good news because, in these cases, one can indeed hope to identify and also evolve structurally diverse compounds having similar activity. The SARI function identifies two subtypes of heterogeneous SARs, 'heterogeneous-relaxed' and 'heterogeneous-constrained' ones. The former category accounts for a situation where continuous and

TABLE 1

SARI profiling of 13 enzyme inhibitor sets

Class	Activity	Continuity score	Discontinuity score	SAR Index	SAR category
LIP	Lipoxygenase inhibitors	0.99	0.04	0.97	Continuous
COX	Cyclooxygenase-2 inhibitors	0.74	0.21	0.77	Continuous
5HT	5-HT reuptake inhibitors	0.56	0.22	0.67	Continuous
ACA	ACAT inhibitors	0.57	0.28	0.64	Continuous
PH4	Phosphodiesterase IV inhibitors	0.66	0.58	0.54	Heterogeneous-relaxed
FXA	Factor Xa inhibitors	0.30	0.27	0.52	Heterogeneous-constrained
ACH	Acetylcholinesterase inhibitors	0.62	0.72	0.45	Heterogeneous-relaxed
FAR	Farnesyl transferase inhibitors	0.58	0.71	0.43	Heterogeneous-relaxed
SQA	Squalene synthase inhibitors	0.79	0.99	0.40	Heterogeneous-relaxed
ELA	Elastase inhibitors	0.36	0.59	0.39	Heterogeneous-constrained
HIV	HIV-1 protease inhibitors	0.12	0.53	0.30	Discontinuous
PH5	Phosphodiesterase V inhibitors	0.08	0.53	0.27	Discontinuous
THR	Thrombin inhibitors	0.08	0.66	0.21	Discontinuous

Inhibitors were assembled from the MDL Drug Data Report (Symyx Software, San Ramon, CA, USA).

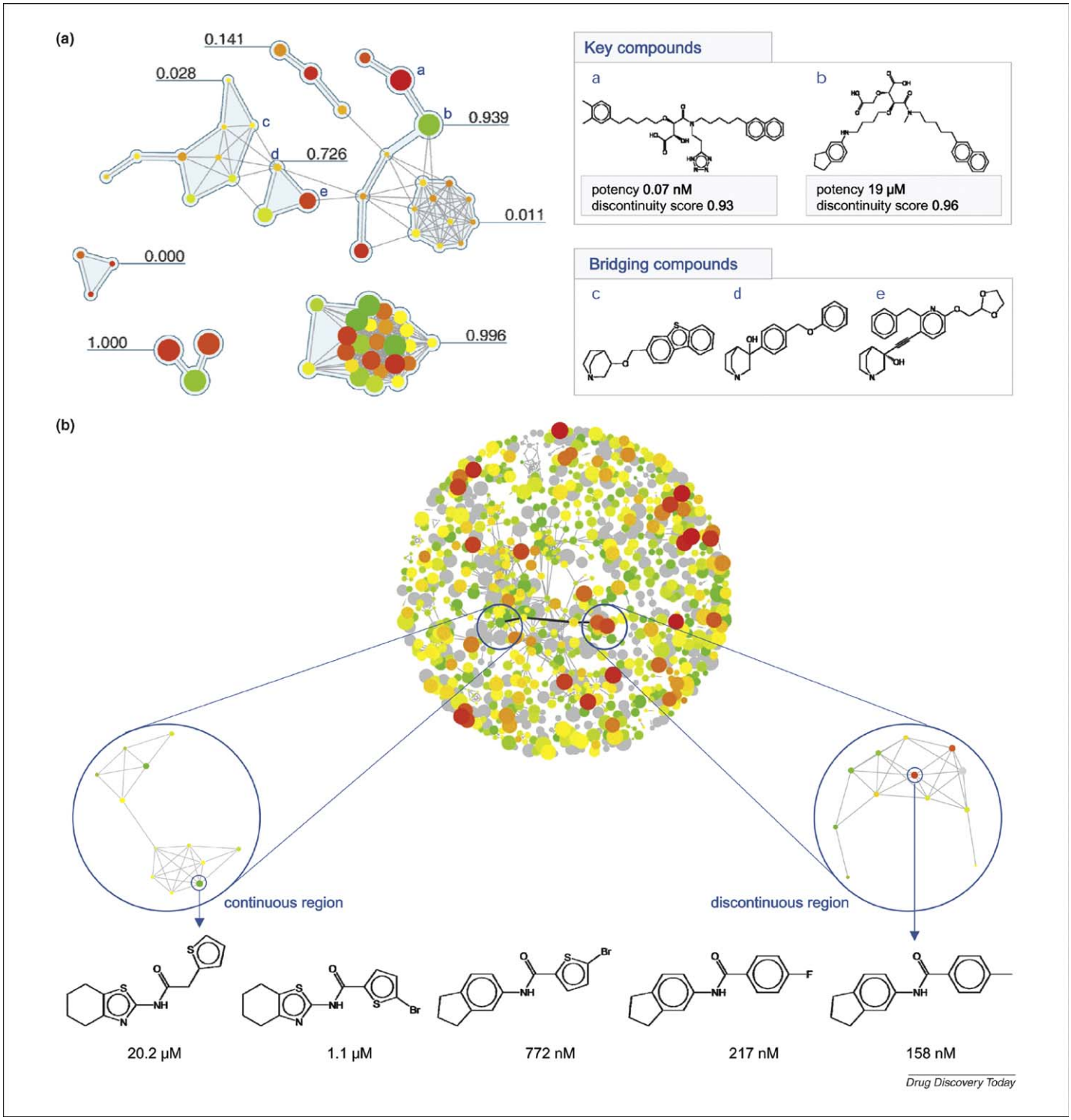


FIGURE 2

Examples of network-like similarity graphs (NSGs). NSGs visualize compounds as nodes and similarity relationships between them as edges (i.e. nodes are connected if their pairwise similarity exceeds a predefined threshold). NSGs combine five different levels of information including pairwise similarity relationships, cluster and potency information, SARI cluster scores and individual SARI compound scores. Nodes are color-coded according to their potency and scaled in size according to the magnitude of discontinuity scores calculated for individual compounds. **(a)** NSG for a set of squalene synthase inhibitors. Compound subsets resulting from global clustering of the dataset are displayed on a gray background and annotated with global discontinuity scores calculated for each cluster. Shown are key compounds, that is 'activity cliff markers', which strongly contribute to local and global SAR discontinuity and, in addition, 'chemical bridges', that is compounds which are similar in structure but connect different local SAR contexts. **(b)** SAR pathway analysis in an NSG of HTS data for compounds increasing NF-κB-expression. The graph includes 842 active and 400 inactive compounds (nodes shown in gray). In the NSG, a pathway is delineated that leads from a weakly active compound located in a region of local SAR continuity to an activity cliff via a series of structurally related compounds with gradually increasing potency. Enlarged portions of the graph show the clusters containing start and end points of the pathway.

discontinuous SAR features coexist within a dataset and the latter for a scenario where SAR continuity is permitted within the boundaries of an activity cliff. Such constraints often correspond to a dominant structural requirement that must be met to ensure effective receptor–ligand interactions, for example, a catalytic cation that must be coordinated within an enzyme's active site. Thus, as long as ligands contain conserved features that are crucial to meet these requirements, other parts of their structure can vary.

To complement global SAR profiling, a local variant of SARI's discontinuity score has been generated that makes it possible to quantify the SAR contribution of each individual active compound [18]. Applying the combination of global and local SARI analysis, activity landscapes can be dissected to identify subsets of active compounds that occur in distinct local SAR contexts and that contribute differently to global SAR characteristics. This type of analysis can be utilized to analyze and interpret complex activity landscapes graphically, beyond the detection of activity cliffs.

SAR network modeling

Molecular network representations have been designed to organize and display similarity and potency relationships systematically within compound datasets [18]. Figure 2a shows a network-like similarity graph (NSG) for a set of squalene synthase inhibitors belonging to the heterogeneous-relaxed SAR category. Multiple activity cliffs and regions of SAR continuity can be readily identified, consistent with the global SAR category. One can even find 'flat' local SARs in the NSG. These are regions that medicinal chemists try to avoid because regardless of structural changes made, there is little effect on biological activity, making potency optimization essentially impossible. In addition to activity cliffs, one also observes 'chemical bridges', that is similar molecules that connect different local SAR contexts. Thus, visual inspection of NSGs enables medicinal chemists to understand, in an intuitive manner, the relationship between local and global SAR characteristics, identify flat SAR regions and activity cliffs and focus on key compounds making significant SAR contributions. Regions of different local SAR character also occur in NSG representations of screening datasets. For large datasets, however, network representations often become too complex for an intuitive visual analysis. Therefore, systematic graph mining procedures are applied to isolate compound pathways that connect different local SAR environments. From such pathways, SAR information can be extracted, as illustrated in Fig. 2b. For example, compound pathways can be identified that consist of series of similar molecules with increasing potency and link continuous SAR regions to activity cliffs. Analysis of such pathways in screening data also supports medicinal chemists in the prioritization of hits for further study.

Summary and outlook

The difficulties associated with exploring SARs continue to be a major bottleneck in early-phase drug discovery. Medicinal chemists are still challenged in most hit-to-lead or lead optimization projects to explore individual SARs on a case-by-case basis. This situation is reminiscent of the problem of navigating the Mississippi delta where many seemingly open paths only lead to a dead-end. The data analysis methods presented herein enable us to study SARs on a large scale, better understand their similarities and differences and assess computational models for the treatment of individual SARs. For the development of these methodologies, the concept of activity landscapes has played a fundamental role. As we have shown, the combination of SAR analysis functions with simple and intuitive graphical representations of activity landscapes helps medicinal chemists to analyze SAR features and focus on key compounds.

By applying analysis functions, our view of SARs widens and we learn more about the limitations of the computational tools we design. For example, perhaps our major objective in navigating activity landscapes should not be the development of computationally complex, and mostly black-box, models but rather of intuitive, data-driven approaches. Focusing on what the data are telling us should help to avoid over-selling the benefits of predictive models. SAR analysis functions do, in principle, not require the availability of more data points than conventional QSAR models, although larger compound sets are often beneficial for the analysis. Importantly, these functions can process raw screening data and are thus already applicable before the hit-to-lead or lead optimization stages of a drug discovery project.

We expect that the approaches described herein will trigger the further development of intuitive computational methods that put SAR analysis on a new level and immediately impact medicinal chemistry programs. Clearly, the ultimate success of tools for SAR exploration will be measured by the degree to which they influence decision-making in medicinal chemistry and lead to measurable successes in the generation of leads and clinical candidates. From our point of view, this is a work in progress.

Competing interests statement

The authors declare no competing financial interests.

Acknowledgements

The authors are grateful to Gerald M. Maggiora and Veerabahu Shanmugasundaram who have been an inspiration for much of the research leading to the development of SAR analysis functions. L.P. is supported by Boehringer Ingelheim and M.W. by the B-IT foundation.

References

1 Sobel, D. (2005) *Longitude*. Penguin Books
2 Lajiness, M.S. *et al.* (2004) Molecular properties that influence oral drug-like behavior. *Curr. Opin. Drug Discov. Dev.* 7, 470–477
3 Lipinski, C.A. *et al.* (1997) Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Deliv. Rev.* 23, 3–25
4 Vieth, M. *et al.* (2004) Characteristic physical properties and structural fragments of marketed oral drugs. *J. Med. Chem.* 47, 224–232
5 Kubinyi, H. (2003) 2D QSAR models: Hansch and Free-Wilson analyses. In *Computational Medicinal Chemistry for Drug Discovery* (Bultinck, P., De Winter, H., Langenaeker, W., Tollenaere, J.P., eds), pp. 539–570, Dekker Inc.
6 Cobb, R.R. and Molony, L.J. (2007) Drug discovery: design and serendipity. In *Principles and Practice of Pharmaceutical Medicine* (Edwards, L.D., Fletcher, A.J., Fox, A.W., Stonier, P.D., eds), pp. 43–50, John Wiley & Sons Ltd
7 Lajiness, M.S. *et al.* (2004) Assessment of the consistency of medicinal chemists in reviewing sets of compounds. *J. Med. Chem.* 47, 4891–4896

8 Kubinyi, H. (2003) Drug research: myths, hype and reality. *Nat. Rev. Drug Discov.* 2, 665–668
9 Johnson, S. (2008) The trouble with QSAR (or how I learned to stop worrying and embrace fallacy). *J. Chem. Inf. Model.* 48, 25–26
10 Maggiora, G.M. (2006) On outliers and activity cliffs – why QSAR often disappoints. *J. Chem. Inf. Model.* 46, 1535–11535
11 Shanmugasundaram, V. and Maggiora, G.M. (2001) Characterizing property and activity landscapes using an information-theoretic approach. *222nd American Chemical Society National Meeting*, Division of Chemical Information Abstract no. 77
12 Peltason, L. and Bajorath, J. (2007) SAR index: quantifying the nature of structure–activity relationships. *J. Med. Chem.* 50, 5571–5578
13 Guha, R. and Van Drie, J.H. (2008) Structure–activity landscape index: identifying and quantifying activity cliffs. *J. Chem. Inf. Model* 48, 646–658
14 Takahashi, H. *et al.* (2006) Discovery and SAR study of novel dihydroquinoline containing glucocorticoid receptor ligands. *Bioorg. Med. Chem. Lett.* 16, 1549–1552
15 Takahashi, H. *et al.* (2007) Discovery and SAR study of novel dihydroquinoline containing glucocorticoid receptor agonists. *Bioorg. Med. Chem. Lett.* 17, 5091–5095
16 Guha, R. and Van Drie, J.H. (2008) Assessing how well a modeling protocol captures a structure–activity landscape. *J. Chem. Inf. Model.* 48, 1716–1728
17 Cramer, R.D. *et al.* (1988) Comparative molecular field analysis (CoMFA). 1. Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc.* 110, 5959–5967
18 Wawer, M. *et al.* (2008) Structure–activity relationship anatomy by network-like similarity graphs and local structure–activity relationship indices. *J. Med. Chem.* 51, 6075–6084