Articles

SAR Index: Quantifying the Nature of Structure−Activity Relationships

Lisa Peltason and Jürgen Bajorath*

Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany

Received May 16, 2007

Structure−activity relationships (SARs) can display very different features. Small chemical modifications of active molecules often dramatically alter biological responses. By contrast, structurally diverse molecules can have similar activity. SARs can also be heterogeneous in nature. For example, for structurally diverse molecules with similar activity, closely related analogs might have significant differences in potency. Given the inherent complexity of SARs, it has been very difficult to estimate SAR characteristics from molecular structure. On the basis of systematic correlation of 2D structural similarity and compound potency, we have developed a function termed "SAR Index" that quantitatively describes the nature of SARs and establishes different SAR categories: continuous, discontinuous, heterogeneous-relaxed, and heterogeneous-constrained. These heterogeneous SAR categories are described for the first time. Given a set of active compounds and their potency values, SAR Index calculations can estimate how likely it is to identify structurally distinct molecules having similar activity.

Introduction

The exploration of structure−activity relationships (SARs) is an important task in medicinal chemistry and drug design.[1] SAR analysis provides a basis for chemical optimization of hits or leads and the identification of novel active molecules. However, it is well-recognized that SAR characteristics often dramatically depend on the types of molecules under study.[2] The magnitude of the response of active compounds to chemical alterations distinguishes many SARs. For example, small chemical modifications can render active molecules completely or nearly inactive or, alternatively, increase their potency.[1] Large-magnitude biological responses to minor chemical changes are characteristic of SARs that are discontinuous in nature.[2] This discontinuity is thought to result from the presence of rugged SAR landscapes and activity cliffs.[3] Other types of SARs are characterized by gradual biological responses to chemical changes.[2] In other words, compounds of increasing structural diversity are often found to display similar activity. This means that the underlying SARs are continuous in nature, which is consistent with the presence of gently sloped SAR landscapes that are reminiscent of rolling hills.[3] Furthermore, molecules can have markedly different core structures, but similar activity. In this case, paths on an activity landscape between different structural classes having similar activity might be complex. However, continuous SARs can also cover such significant degrees of structural diversity. Importantly, continuous SARs are consistent with the well-known similarity-property principle[4] that provides the conceptual basis for molecular similarity-based virtual screening[2,5] and the identification of different structural motifs having similar functions.[5] While discontinuous SARs are typically exploited in lead optimization, continuous SARs provide the basis for the identification of structurally diverse hits in virtual or experimental compound screening.[2] When analyzing SAR characteristics, it must also be considered that the nature of activity landscapes is generally greatly influenced by chosen molecular representations, chemical reference spaces, and similarity measures.

Recent studies have shown that the presence of such distinct SAR characteristics is not mutually exclusive for series of active compounds.[6] Through the systematic correlation of two- and three-dimensional similarity and potency of active compounds taken from X-ray structures of complexes with their target proteins, it could be established that continuous and discontinuous SARs often coexist in different series of active compounds.[6] These studies confirmed an earlier proposal that many small molecule SARs should be heterogeneous in nature and characterized by SAR landscapes where activity cliffs are separated by gently sloped or flat regions.[2] Figure 1 shows an example of kinase inhibitors that illustrates heterogeneous SAR characteristics. Molecules with different core structures or scaffolds are strongly potent, but closely related analogs of each scaffold have dramatically reduced potency. Clearly, such relationships greatly complicate the identification, design, or optimization of active compounds.

The complex nature of small molecule SARs has made it difficult to systematically study or classify these relationships. Typically, compound classes and SARs are investigated on a case-by-case basis. Moreover, while qualitative evidence concerning different types of SARs is accumulating, there is currently no quantitative measure available to capture and represent SAR characteristics. With the introduction of the SAR Index (SARI), we have attempted to put the characterization and comparison of SARs on a quantitative basis. The SARI is calculated from two individual scores that capture the two-dimensional (2D) structural diversity and potency distribution within a set of active compounds and range from 0 to 1. On the basis of individual and combined SARI scores, we can distinguish between different categories of SARs. In this study, we introduce the SARI function and quantitatively characterize SARs for 16 different sets of enzyme inhibitors.

* To whom correspondence should be addressed. Tel.: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.

10.1021/jm0705713 CCC: $37.00
Published on Web 09/29/2007

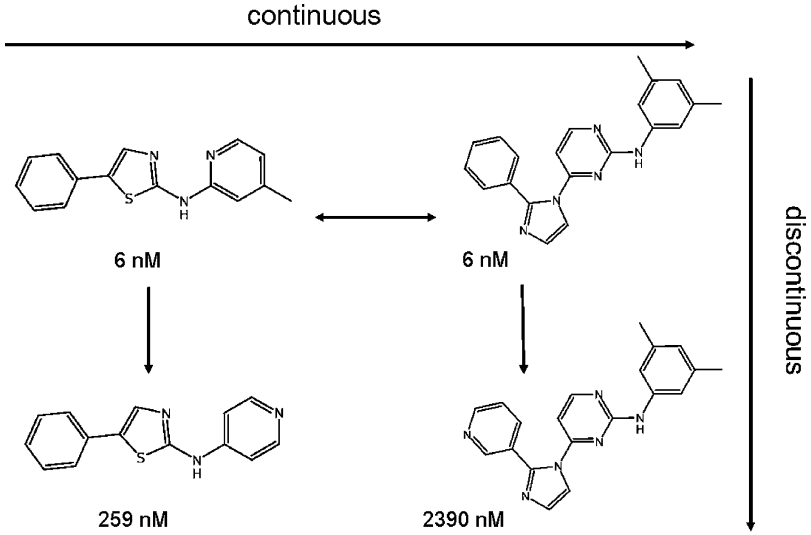

Figure 1. Example of a heterogeneous SAR. Four inhibitors of vascular endothelial growth factor receptor (VEGFR-2) tyrosine kinase are shown. Two inhibitors have different core structures but are highly potent and thus are indicative of a continuous SAR. However, a closely related analog of each inhibitor has 2−3 orders of magnitude potency lower, which exemplifies discontinuous SARs. Thus, this set of kinase inhibitors is characterized by a heterogeneous SAR. The example is adapted from ref 2.

SAR Index

First we describe the conceptual framework of SARI. We generally distinguish between three SAR categories: continuous, discontinuous, and heterogeneous SARs. The SARI function is designed to quantify these SAR characteristics for given sets of compounds active against specific targets. Therefore, SARI is constituted by two separately calculated scores that assess intra-class diversity and also the potency differences of similar ligands as the major determinants of SAR characteristics. The "continuity" score ($score_{cont}$) measures potency-weighted structural diversity within a class of active compounds. High continuity scores reflect the presence of structurally diverse molecules having comparable potency, which is a major characteristic of continuous SARs. The "discontinuity" score ($score_{disc}$) determines average potency differences for pairs of similar ligands, which reveals the presence of activity cliffs as a major determinant of discontinuous SARs.

The continuity score is derived from the potency-weighted mean of pairwise ligand similarity. Ligand pairs are weighted by the magnitude of their potency and the reciprocal of their potency difference. Accordingly, pairs of ligands with high potency but low potency differences contribute more to the continuity score than ligand pairs with low potency and high potency differences. This weighting scheme takes into account that continuity of SARs is mostly reflected by the presence of similarly potent inhibitors of increasing structural diversity. The potency-weighted mean of ligand similarity is transformed into a diversity measure by subtraction from 1.

The discontinuity score measures the average potency differences for pairs of ligands multiplied by pairwise ligand similarity. Here only ligand pairs are considered that reach a predefined similarity threshold value because discontinuity of SARs is largely reflected by the presence of similar compounds with significant differences in potency. High discontinuity scores are therefore indicative of discontinuous SAR characteristics.

The continuity and discontinuity scores are normalized to yield values between 0 and 1 and the final SARI is defined as the mean of the transformed scores. A high continuity score is an indicator for continuous SARs, whereas a high discontinuity score means the opposite. Consequently, the discontinuity score is subtracted from 1 to obtain a complementary score value

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

Like the individual scores, the SARI yields values between 0 and 1. Low SARI values indicate discontinuous SARs, high SARI values indicate continuous, and intermediate values indicate heterogeneous SARs that combine continuous and discontinuous elements. The mathematical derivation of the SARI function and further details are presented in Methods.

SAR Characteristics of Different Classes of Enzyme Inhibitors. For our analysis, we have searched for compound activity classes that cover a wide range of targets and have varying degrees of structural diversity and significantly different potency distributions. Given these criteria, we have selected 16 sets of enzyme inhibitors consisting of between 9 and 33 compounds, as summarized in Table 1. As can be seen, most activity classes have a large intra-class diversity spread and differ significantly in their potency distribution. This is further illustrated in Figure 2, which reports the correlation between 2D similarity and compound potency differences for six selected classes on the basis of pairwise compound comparisons and confirms the presence of significant differences between these classes.

We then calculated SAR Indices for the enzyme inhibitor sets and found that they covered a broad spectrum of SAR characteristics. The calculated continuity, discontinuity, and SARI scores for the enzyme inhibitor sets are reported in Table 2 and differ substantially. Continuity scores ranged from essentially 0 to 0.82, discontinuity scores from 0.03 to 0.93, and the resulting SARI scores from 0.15 to 0.89. Two sets of inhibitors produced SARI scores smaller than 0.2 and two other SARI scores of 0.8 or greater. A total of 8 of our 16 classes fell into an intermediate scoring range between 0.40 and 0.55.

On the basis of our calculations, we can essentially distinguish between four categories of SARs that are present in our compound reference sets: (i) high SARI scores (high continuity, low discontinuity scores) are indicative of continuous SARs (prototypic example: factor Xa); (ii) low SARI scores (low continuity, high discontinuity scores) are indicative of discon-

Table 1. Enzyme Inhibitor Classes and Their Similarity and Potency Distributions[a]

target	E.C. #	cmpd	MACCS Tc			potency (K_i values, nM)		
			min	max	mean	min	max	mean
acetylcholine esterase	3.1.1.7	19	0.22	1	0.46	0.13	8900	1210
adenosine deaminase	3.5.4.4	19	0.34	1	0.67	0.0001	9000	1254
carbonic anhydrase II	4.2.1.1	27	0.07	1	0.59	0.03	125 000	4669
coagulation factor Xa	3.4.21.6	16	0.24	1	0.50	0.007	131	24
cyclin-dependent kinase 2	2.7.11.22	27	0.25	0.99	0.49	3	38 000	6017
cyclooxygenase 2	1.14.99.1	21	0.21	0.96	0.48	0.09	3380	396
dihydrofolate reductase	1.5.1.3	23	0.34	0.84	0.54	0.1	19 500	1472
elastase	3.4.21.36	14	0.34	0.91	0.52	0.46	890 000	120 512
peptidylprolyl isomerase (FKBP-12)	5.2.1.8	14	0.09	0.99	0.54	0.2	500 000	53 675
poly(ADP−ribose) polymerase	2.4.2.30	23	0.26	0.82	0.47	5	35 000	1615
protein-tyrosine phosphatase 1b	3.1.3.48	22	0.13	0.91	0.49	1.8	63 000	7635
ribonuclease A	3.1.27.5	9	0.76	0.99	0.87	27	82 000	12 821
thrombin	3.4.21.5	28	0.23	1	0.49	0.0014	85 000	9018
thromboxane synthase	5.3.99.5	23	0.21	1	0.48	0.8	33 000	2144
thymidylate synthase	2.1.1.45	18	0.29	0.98	0.66	2	36 000	2260
trypsin	3.4.21.4	33	0.14	1	0.49	5.2	32 500 000	1 825 334

[a] Enzyme inhibitor sets were collected from various sources (see Methods).

tinuous SARs (e.g., adenosine deaminase); (iii) intermediate SARI scores (~0.50) produced by high continuity and discontinuity scores are an indicator of heterogeneous SARs (e.g., thromboxane synthase); and (iv) intermediate SARI scores produced by low continuity and discontinuity scores indicate (another kind of) heterogeneous SARs (e.g., thymidylate synthase).

SAR categories (iii) and (iv) are both characterized by intermediate SARI scores pointing at heterogeneous SARs but are distinguished by the magnitude of continuity and discontinuity scores (i.e., high/high versus low/low). As will be discussed in detail below, the relationship between the different scores distinguishes between two different types of heterogeneous SARs: one is characterized by coexisting different continuous and discontinuous SARs, whereas the other displays continuous SAR characteristics within the constraints presented by an activity cliff.

Quantitative Description of Representative SARs. In the following, we discuss the results of our SARI analysis for selected inhibitor classes that are representative of the four SAR categories described above.

Factor Xa. The similarity versus potency distribution of factor Xa inhibitors in Figure 2a reveals a wide distribution of potent inhibitor pairs over the Tanimoto coefficient (Tc) range 0.2 to 1. The factor Xa inhibitor set produced a high continuity (0.71) and a low discontinuity (0.12) score, resulting in a SARI score of 0.80, which is a clear indicator of a continuous SAR. Accordingly, factor Xa inhibitors represent a class of high structural diversity among highly potent molecules (the potency-weighted mean of pairwise similarity is 0.51) but potency differences in pairs of compounds with MACCS Tc > 0.6 are low. Figure 3 shows increasingly diverse factor Xa inhibitors, all of which have nanomolar potency. This figure qualitatively illustrates the presence of a continuous SAR, consistent with the conclusion drawn from continuity, discontinuity, and SARI scoring.

Ribonuclease A. In marked contrast to factor Xa, the ribonuclease A inhibitor set produced a continuity score close to zero (0.004), a discontinuity score of 0.68, and a SARI score of 0.16, which clearly indicates the presence of a discontinuous SAR. Inhibitors within this set have highly similar 2D structures (potency-weighted mean of pairwise similarity: 0.88), but even pairs of ligands with MACCS Tc similarity between 0.75 and 0.9 have potency differences of up to 2.5 orders of magnitude (Figure 2b). Thus, in this case, the SAR is dominated by the influence of an activity cliff, which can be well rationalized taking into account that highly potent ribonuclease A inhibitors are nucleotide analogs whose activity depends on the presence of specific phosphate groups. This is illustrated in Figure 4, which shows two inhibitors with significantly different potency due to the presence or absence of two phosphate groups. Again, SARI scores are found to quantitatively reflect qualitative characteristics of a specific SAR type; in this case, a discontinuous SAR.

Adenosine Deaminase. With continuity, discontinuity, and SARI scores of 0.15, 0.85, and 0.15, respectively, inhibitors of adenosine deaminase are also related by a discontinuous SAR. Within this class, many potent inhibitors are structurally similar, but ligand pairs with low similarity (Tc < 0.5) have only moderate potency in the micromolar range. Figure 2c illustrates these trends. The SAR is dominated by an activity cliff that results from the requirement to coordinate a zinc cation in the enzyme's active site. Potent inhibitors contribute a hydroxyl group in a defined spatial position to the coordination sphere. Figure 5 illustrates the influence of this activity cliff. Here closely related adenosine deaminase inhibitors have dramatic differences in potency dependent on whether or not the coordinating hydroxyl group is present. SARI calculations correctly characterize this discontinuous SAR only on the basis of 2D structural information and potency values of inhibitors.

Thromboxane Synthase. In contrast to the three classes discussed above, the thromboxane synthase inhibitor set produced an intermediate SARI score of close to 0.5 (0.46), which resulted from high continuity (0.82) and high discontinuity (0.89) scores. These high individual scores indicate that continuous and discontinuous elements coexist within the SAR landscape of this enzyme. SAR continuity can be appreciated in Figure 2d where highly potent inhibitor pairs concentrate at lower 2D similarity levels (potency-weighted mean of similarity: 0.47). However, there is also a significant number of similar structures having large differences in potency, which reflects SAR discontinuity. Figure 6 shows a series of thromboxane synthase inhibitors in which 2D similarity and potency are gradually decreasing, which clearly represents a continuous SAR component. In addition, Figure 7 shows a structurally diverse subset of inhibitors that have high potency in the nanomolar range, thus representing another continuous SAR element. By contrast, Figure 8 depicts two closely related analogs belonging to another

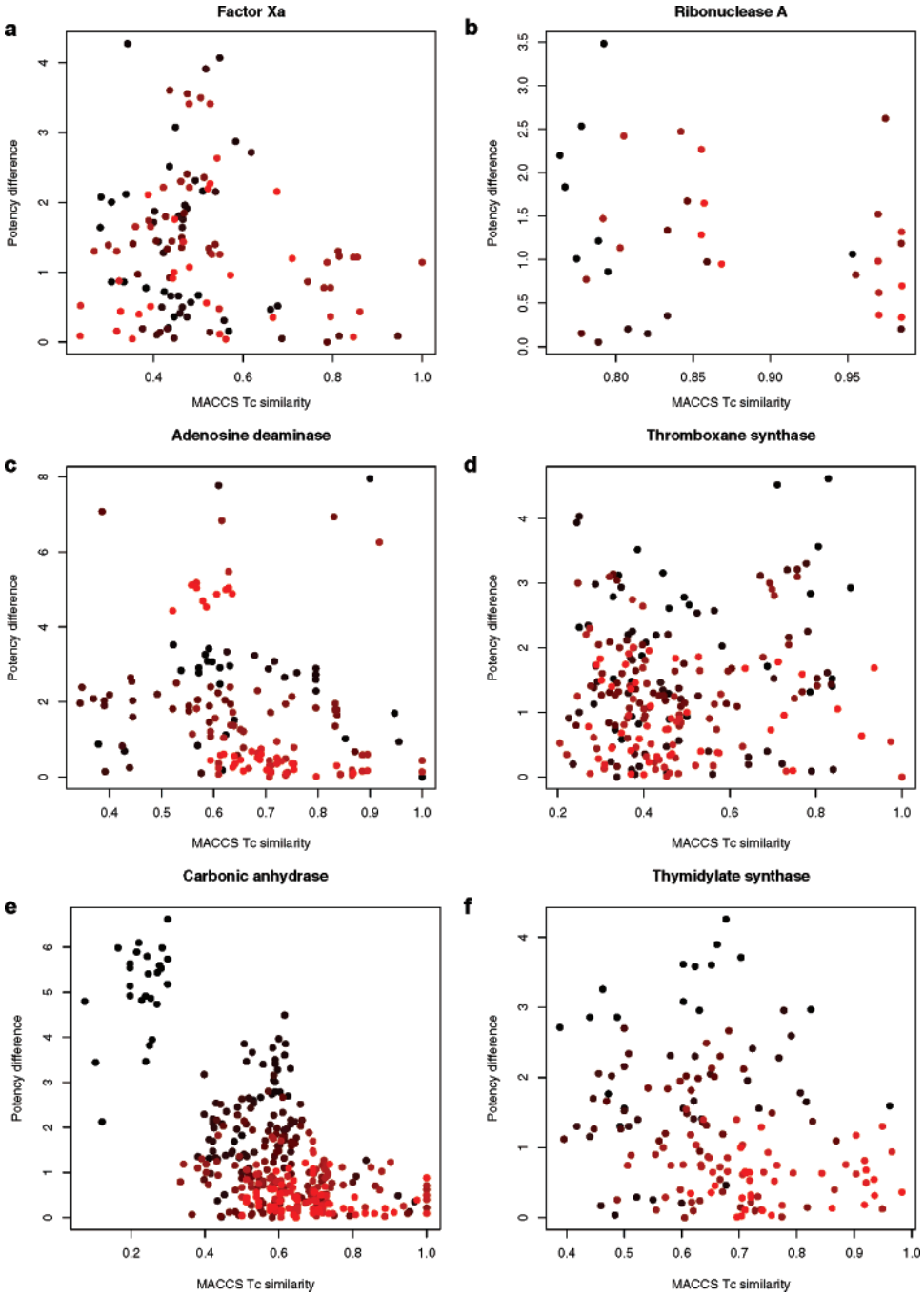

Figure 2. Potency differences versus 2D similarity of enzyme inhibitors. Each data point represents a pairwise comparison of inhibitors within an activity class. Data points (i.e., pairs of inhibitors) are color-coded according to potency measured by the sum of their pK_i values using a continuous spectrum from black (lowest potency) to red (highest potency). Distributions are shown for six sets of enzyme inhibitors that represent different types of SARs, as discussed in the text: (a) factor Xa, (b) ribonuclease A, (c) adenosine deaminase, (d) thromboxane synthase, (e) carbonic anhydrase, (f) thymidylate synthase.

series of thromboxane synthase inhibitors that have significantly different potencies, which is characteristic of a discontinuous SAR.

Thus, in the case of thromboxane synthase, different continuous and discontinuous SAR components mutually coexist. The enzyme is permissive to different types of small molecule SARs. Therefore, we term this SAR type "heterogeneous-relaxed". As stated above, due to its heterogeneous nature, this SAR is characterized by an intermediate SARI score. However, its distinguishing features are high continuity and high discontinuity scores.

Carbonic Anhydrase. Inhibitors of carbonic anhydrase also represent a heterogeneous SAR. The SARI score is 0.61 and Figure 2e shows that inhibitor pairs with high potency span a MACCS Tc range of about 0.5 to 1. However, in marked contrast to thromboxane synthase, the intermediate SARI score of carbonic anhydrase inhibitors is the result of low continuity (0.30) and low discontinuity (0.08) scores. Although perhaps puzzling at a first glance, this SAR phenotype can be well rationalized: it is characterized by SAR continuity within the boundaries of a structural constraint. The major determinant for carbonic anhydrase inhibitory activity is the presence of a sulfonamide group that complexes a zinc cation in the enzyme's active site, similar to the situation of adenosine deaminase discussed above. However, in contrast to adenosine deaminase, carbonic anhydrase permits significant scaffold diversity among

Table 2. SAR Indices for Different Classes of Enzyme Inhibitors.

target	SAR characteristics		
	continuity score	discontinuity score	SARI
acetylcholine esterase	0.82	0.93	0.45
adenosine deaminase	0.15	0.85	0.15
carbonic anhydrase II	0.30	0.08	0.61
coagulation factor Xa	0.71	0.12	0.80
cyclin-dependent kinase 2	0.74	0.36	0.69
cyclooxygenase 2	0.79	0.69	0.55
dihydrofolate reductase	0.59	0.67	0.46
elastase	0.64	0.38	0.63
peptidylprolyl isomerase (FKBP-12)	0.17	0.26	0.45
poly(ADP−ribose) polymerase	0.82	0.03	0.89
protein-tyrosine phosphatase 1b	0.75	0.44	0.66
ribonuclease A	0.004	0.68	0.16
thrombin	0.71	0.92	0.40
thromboxane synthase	0.82	0.89	0.46
thymidylate synthase	0.16	0.33	0.41
trypsin	0.37	0.42	0.47

inhibitors as long as the sulfonamide constraint is met. Moreover, similar compounds always display comparable potency. This is illustrated in Figure 9, which reveals a continuous SAR for diverse sulfonamide-containing inhibitors. If the sulfonamide group is present, inhibitors have high potency in the low nanomolar range; if it is absent, the potency is reduced by 4−6 orders of magnitude (see upper left region in Figure 2e).

The heterogeneous SAR exemplified by the carbonic anhydrase inhibitor set is distinct from the heterogeneous SAR phenotype presented by thromboxane synthase inhibitors. In the latter case, different continuous and discontinuous SARs coexist, whereas in the case of carbonic anhydrase, a continuous SAR is observed within the limits of a structural constraint. Therefore, we term the SAR type exemplified by carbonic anhydrase inhibitors "heterogeneous-constrained". SARI analysis clearly distinguishes these subtypes of heterogeneous SARs. In contrast to heterogeneous-relaxed SARs that are characterized by high continuity and discontinuity scores, the distinguishing features of heterogeneous-constrained SARs are low continuity and low discontinuity scores.

Thymidylate Synthase. The thymidylate synthase inhibitor set presents another example of a heterogeneous-constrained SAR. For this set, the continuity, discontinuity, and SARI scores are 0.16, 0.33, and 0.41, respectively. Figure 2f reveals a similarity versus potency distribution where inhibitor pairs with high potency concentrate within a MACCS Tc range of about 0.6 to 1, similar to carbonic anhydrase. The presence of a structural constraint indicated by the low continuity score is illustrated by the fact that all inhibitors are quinazolone or pyrimidine derivatives. However, as long as these core structures are present, structural variations among inhibitors can occur and are accompanied by systematic changes in potency. Figure 10 illustrates this SAR continuity; gradual structural departures from a potent inhibitor (on the left in Figure 10) result in substantial decreases in potency, and these changes are of larger magnitude than has been observed for carbonic anhydrase. This suggests that there is less SAR continuity within the boundaries of the structural constraint in the case of thymidylate synthase compared to carbonic anhydrase. This observation is consistent with the fact that the thymidylate synthase inhibitors produced a lower SARI score (0.41 vs 0.61) that is shifted more toward the low scoring range characteristic for discontinuous SARs. These findings indicate that the SARI scoring scheme provides a sensitive measure to estimate the balance between continuous and discontinuous components of heterogeneous SARs.

Methods

Compound Data and Similarity. Different sets of enzyme inhibitors with reported potency values were taken from PDBbind,[7,8] the Molecular Drug Data Report (MDDR),[9] or literature sources. For pairs of ligands, potency differences were determined as the absolute difference between their pK_i or pIC_{50} values. Two-dimensional structural similarity of inhibitors was calculated using the MACCS fingerprint representations[10] and the Tc.[11] The MACCS fingerprint is a publicly available structural fragment-type molecular

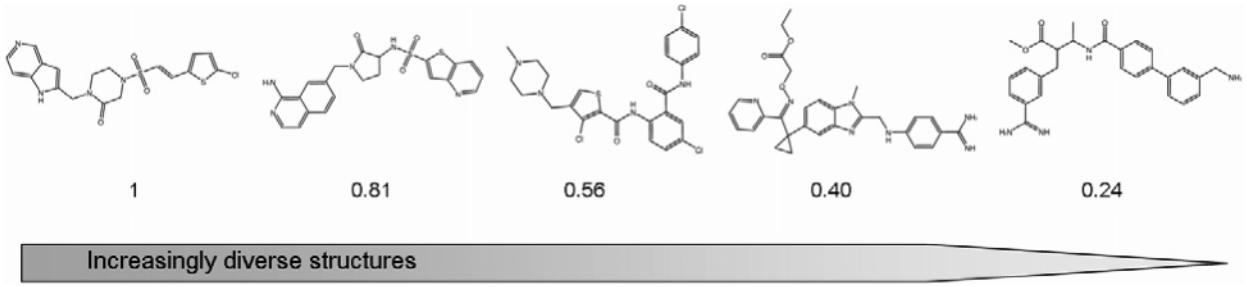

Figure 3. Spectrum of potent factor Xa inhibitors with increasingly diverse structures having comparable potency in the nanomolar range. MACCS Tc values are reported for pairwise compound comparisons using the compound on the left side as the reference molecule.

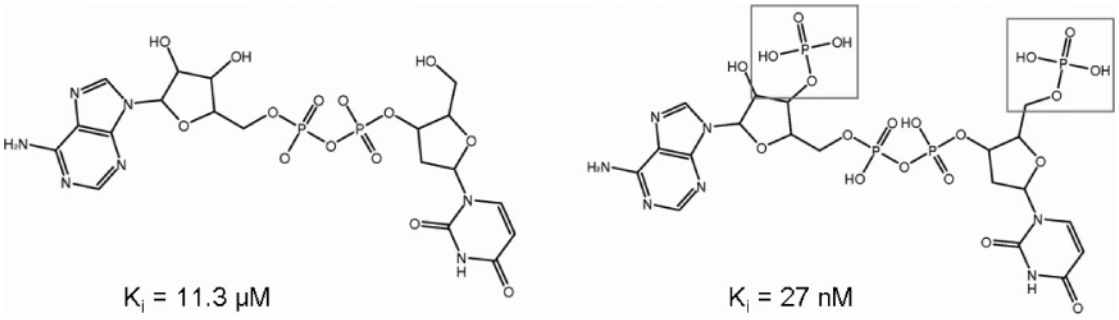

Figure 4. A pair of ribonuclease A inhibitors with significantly different potency that are only distinguished by two phosphate groups.

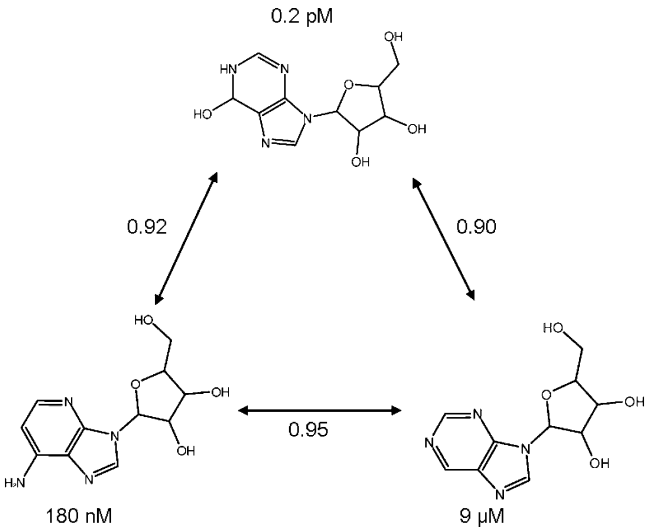

Figure 5. Adenosine deaminase inhibitors. Closely related analogs with high potency differences. MACCS Tc values are reported for pairwise comparisons.

descriptor consisting of 166 bit positions. Each bit accounts for a specific structural fragment in the molecular graph representation. For each compound, MACCS was computed using the Molecular Operating Environment (MOE).[12] The conventional Tc served as a measure of bit string overlap. It counts the number of bits common to two binary fingerprints with respect to the total number of bits that are set on in each fingerprint. The Tc for two binary fingerprint representations *A* and *B* is calculated as follows

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

where 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. Given this formulation, identical fingerprints have a Tc value of 1, whereas nonoverlapping fingerprints are assigned a Tc value of 0. For each compound set, systematic pairwise comparisons of potency differences and Tanimoto similarity were carried out and graphically analyzed. These correlation plots are conceptually similar to structure−activity similarity maps.[13]

Derivation of the SARI. The SARI consists of two separately calculated scores. The continuity score ($\mathrm{score}_{\mathrm{cont}}$) measures the potency-weighted structural diversity within a given compound set,

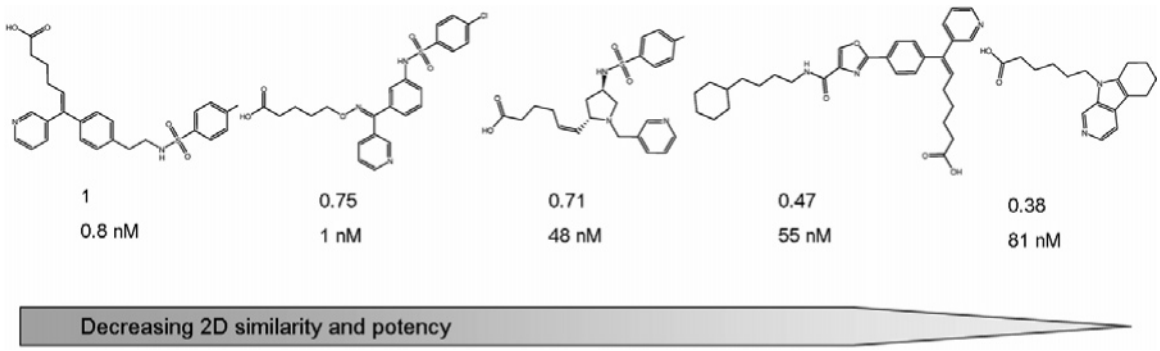

Figure 6. Continuous relationship between 2D similarity and potency within a series of thromboxane synthase inhibitors. The representation is according to Figure 3.

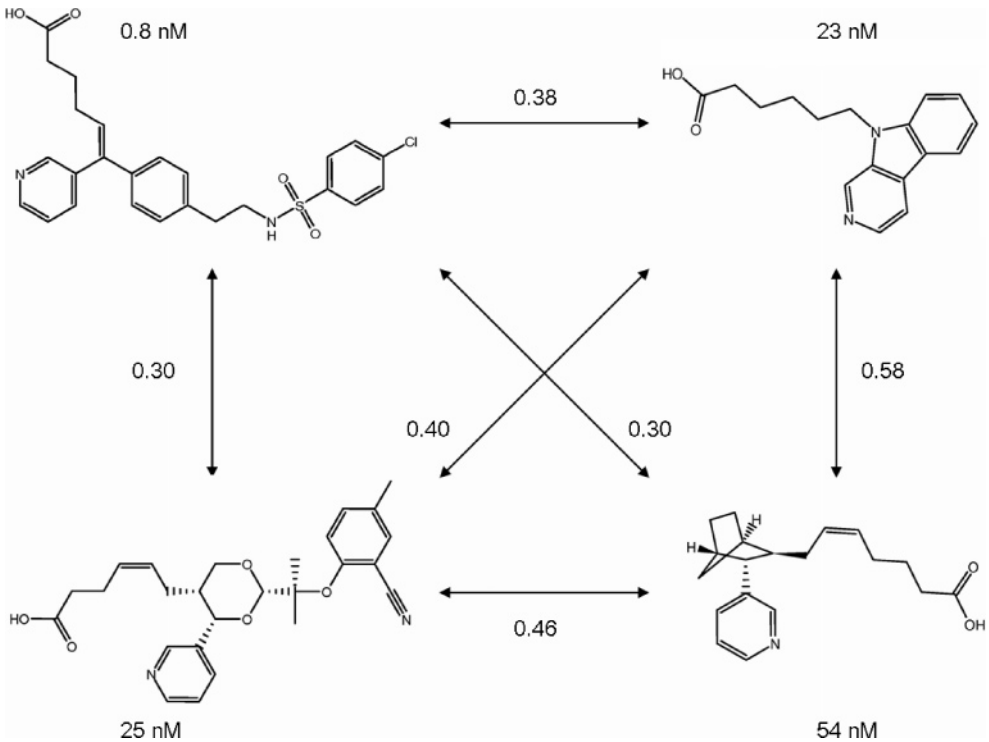

Figure 7. Structurally diverse set of thromboxane synthase inhibitors with comparably high potency.

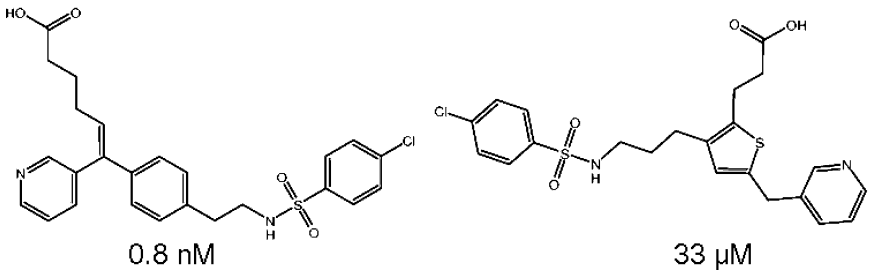

Figure 8. Thromboxane synthase inhibitors with high structural similarity but dramatically different potency.

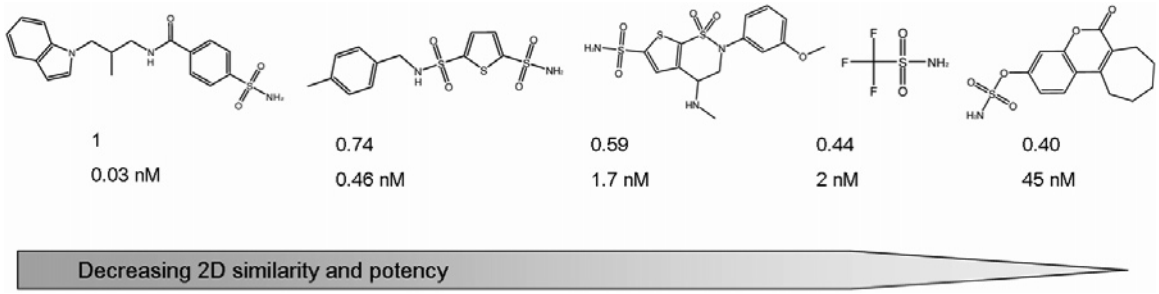

Figure 9. Carbonic anhydrase inhibitors. Increasingly diverse analogs display gradual changes in potency.

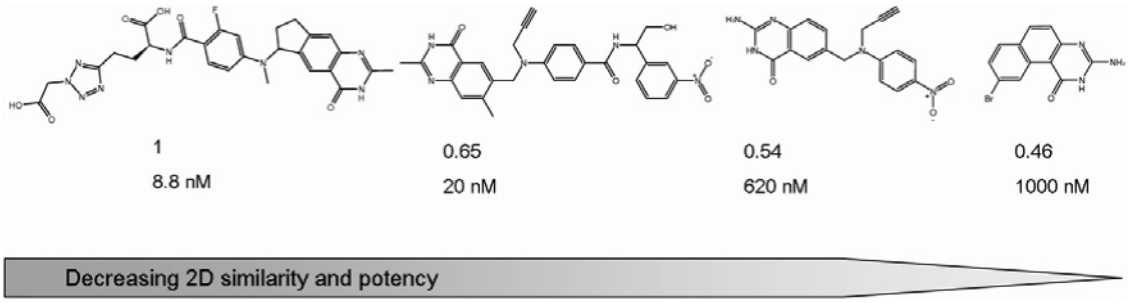

Figure 10. Thymidylate synthase inhibitors. Increasingly diverse quinazolones have decreasing potency.

whereas the discontinuity score ($score_{disc}$) captures the average potency difference among similar compound pairs, as defined below.

The continuity score is derived from the weighted mean of pairwise compound similarity. For each ligand pair, the weights combine the magnitude of their potency and also the potency difference. The "raw" score is defined as follows where the weight

$$\mathrm{raw}_{\mathrm{cont}} = 1 - \frac{\sum_{\text{ligands } i>j} w_{ij}\mathrm{sim}(i,j)}{\sum_{\text{ligands } i>j} w_{ij}}$$

for each ligand pair (i,j) is set to

$$w_{ij} = \frac{\mathrm{pot}(i) \times \mathrm{pot}(j)}{1 + |\mathrm{pot}(i) - \mathrm{pot}(j)|}$$

In this formula, pot(i) gives the potency value of compound i as a pK_i (or pIC_{50}) value and sim(i,j) refers to the MACCS Tanimoto similarity between compounds i and j.

The "raw" discontinuity score is defined as the average potency difference between ligands with MACCS Tc greater than 0.6 multiplied by the pairwise ligand similarity

$$\mathrm{raw}_{\mathrm{disc}} = \frac{\sum_{\{i,j|\mathrm{sim}(i,j)>0.6,\, i>j\}} |\mathrm{pot}(i) - \mathrm{pot}(j)| \times \mathrm{sim}(i,j)}{|\{i,j|\mathrm{sim}(i,j) > 0.6,\, i > j\}|}$$

To form a pair for the calculation of the discontinuity score, two ligands are considered similar if their MACCS Tc value is greater than 0.6. This is a relatively "soft" similarity threshold value. The similarity threshold value is set only for the calculation of the discontinuity score because it focuses on potency differences between similar compounds as an indicator of activity cliffs. Increasing the similarity threshold value did not significantly affect score calculations. Multiplication by ligand similarity emphasizes potency differences among highly similar compounds and produces higher discontinuity scores for activity classes that contain such characteristics.

For ease of comparison, raw scores are converted to Z-scores using the sample mean ($\overline{\mathrm{raw}}$) and standard deviation (sd(raw)) of the scores of a set of reference classes

$$\mathrm{zscore}_{\mathrm{cont}} = \frac{\mathrm{raw}_{\mathrm{cont}} - \overline{\mathrm{raw}_{\mathrm{cont}}}}{\mathrm{sd}(\mathrm{raw}_{\mathrm{cont}})}$$

$$\mathrm{zscore}_{\mathrm{disc}} = \frac{\mathrm{raw}_{\mathrm{disc}} - \overline{\mathrm{raw}_{\mathrm{disc}}}}{\mathrm{sd}(\mathrm{raw}_{\mathrm{disc}})}$$

To calculate Z-scores, we used the 16 compound activity classes assembled for our study as the reference set. We found that using larger reference sets with other activity classes did not measurably

change the results of our analysis. Thus, the classes studied here were sufficient to calculate statistically sound Z-scores. The scores are then transformed into the value range [0,1] by calculating the cumulative probability distribution for each score under the assumption of a normal distribution, which yields the final continuity, discontinuity, and SARI scores

$$\text{score}_{\text{cont}} = \Phi(\text{zscore}_{\text{cont}}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\text{zscore}_{\text{cont}}} \exp\left(-\frac{1}{2}x^2\right)\text{d}x$$

$$\text{score}_{\text{disc}} = \Phi(\text{zscore}_{\text{disc}}) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\text{zscore}_{\text{disc}}} \exp\left(-\frac{1}{2}x^2\right)\text{d}x$$

Concluding Remarks

With the SARI and its individual score components, we have introduced an approach to quantitatively describe the nature of SARs. To characterize SARs, we only require 2D representations and potency values for sets of compounds sharing the same biological activity. The results of our calculations on different sets of enzyme inhibitors show that SARI calculations can quantitatively distinguish between different SAR categories. Thus, these calculations are useful to classify SARs on a large scale. The majority of compound classes tested in our analysis produced intermediate SARI scores that are indicative of heterogeneous SARs. These findings are consistent with earlier proposals that many small molecule SARs should be heterogeneous in nature.[2] Otherwise, the ability to identify diverse structural motifs having similar activity in compound screening campaigns and subsequently optimize these compounds would be difficult to rationalize. Furthermore, on the basis of our analysis, we have been able to further divide heterogeneous SARs into two previously unobserved categories (heterogeneous-relaxed and heterogeneous-constrained) that are distinguished by the magnitude of continuity and discontinuity scores and have different characteristics. In addition to SAR classification, the SARI approach also has significant potential for other practical applications. For example, given a set of active compounds and their potencies, SARI calculations make it possible to estimate how likely it might be to identify structurally diverse compounds having similar activity: the larger the SARI score, the higher the probability to do so.

References

(1) Kubinyi, H. Similarity and dissimilarity. A medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 225−252.
(2) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: Foundations, limitations, and novel approaches. *Drug Discovery Today* **2007**, *12*, 225−233.
(3) Maggiora, G. M. On outliers and activity cliffs—Why QSAR often disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535−1535.
(4) Johnson, M., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; John Wiley & Sons: New York, 1990.
(5) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.
(6) Peltason, L.; Bajorath, J. Molecular similarity analysis uncovers heterogeneous structure−activity relationships and variable activity landscapes. *Chem. Biol.* **2007**, *14*, 489−497.
(7) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein−ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977−2980.
(8) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; and Wang, S. The PDBbind database: Methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111−4119.
(9) *Molecular Drug Data Report (MDDR)*. Elsevier MDL, San Leandro, CA, http://www.mdl.com (accessed Sept 1, 2006).
(10) *MACCS structural keys*. Elsevier MDL, San Leandro, CA, http://www.mdl.com (accessed Sept 1, 2006).
(11) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183−4199.
(12) *Molecular Operating Environment (MOE)*, version 2005.06; Chemical Computing Group, Inc.: 1255 University Street, Montreal, Quebec, Canada, H3B 3X3 (http://www.chemcomp.com (accessed Nov 1, 2005)).
(13) Shanmugasundaram, V.; Maggiora, G. M. *Characterizing property and activity landscapes using an information-theoretic approach*, 222nd American Chemical Society National Meeting, Division of Chemical Information; American Chemical Society: Washington, DC, 2001; Abstract no. 77.

JM0705713