Structural Interpretation of Activity Cliffs Revealed by Systematic Analysis of Structure−Activity Relationships in Analog Series

Mihiret T. Sisay,[†,‡,§] Lisa Peltason,[†,§] and Jürgen Bajorath[*,†]

Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany, and Pharmazeutisches Institut, Pharmazeutische Chemie I, Rheinische Friedrich-Wilhelms-Universität, An der Immenburg 4, D-53121 Bonn, Germany

Discontinuity in structure−activity relationships (SARs) is caused by so-called activity cliffs and represents one of the major caveats in SAR modeling and lead optimization. At activity cliffs, small structural modifications of compounds lead to substantial differences in potency that are essentially unpredictable using quantitative structure−activity relationship (QSAR) methods. In order to better understand SAR discontinuity at the molecular level of detail, we have analyzed different compound series in combinatorial analog graphs and determined substitution patterns that introduce activity cliffs of varying magnitude. So identified SAR determinants were then analyzed on the basis of complex crystal structures to enable a structural interpretation of SAR discontinuity and underlying activity cliffs. In some instances, SAR discontinuity detected within analog series could be well rationalized on the basis of structural data, whereas in others a structural explanation was not possible. This reflects the intrinsic complexity of small molecule SARs and suggests that the analysis of short-range receptor−ligand interactions seen in X-ray structures is insufficient to comprehensively account for SAR discontinuity. However, in other cases, SAR information extracted from ligands was incomplete but could be deduced taking X-ray data into account. Thus, taken together, these findings illustrate the complementarity of ligand-based SAR analysis and structural information.

INTRODUCTION

In hit-to-lead and lead optimization projects, active compounds are subjected to chemical modification, and series of analogs are generated from which SAR information is extracted. In analog design and exploration, one typically attempts to identify substitution sites where R-group variations lead to improved potency (and other desired compound characteristics) and SAR patterns that can be rationalized and ultimately used to predict highly potent compounds.[1] However, the derivation of SAR rules and guidelines for optimization is often severely compromised by abrupt changes in the biological response to minor chemical modifications of active compounds. This situation is typically attributed to the presence of an activity cliff[2] or a region of SAR discontinuity.[1] Both of these terms are derived from the intuitive concept of an activity landscape[1−3] that describes activity responses to positional changes in biologically relevant chemical space. At activity cliffs, small changes in structure, corresponding to small steps in chemical space, lead to significant changes in the activity/potency hypersurface. Multiple activity cliffs can be present in the activity landscape shaped by a compound series,[1] and each of these cliffs gives rise to local SAR discontinuity observed for a compound subset. Accordingly, the terms activity cliff and SAR discontinuity are conceptually linked and can be

used interchangeably. The magnitude of activity cliffs is generally influenced by the chosen chemical reference space and molecular representation. The same applies to the location of different activity islands, i.e., small regions in chemical space that are enriched with compounds sharing a specific biological activity.

SAR analysis functions,[3] such as the SAR Index (SARI),[4] have been designed to quantitatively account for SAR discontinuity (and also continuity) within compound data sets. In principle, a compound set is characterized by low SAR discontinuity if it consists of moderately similar or even structurally diverse compounds having only relatively small differences in potency and, in contrast, by high discontinuity, if it contains very similar compounds with dramatic potency differences.

A systematic study of such SAR patterns within analog series is generally complicated because SARs are typically heterogeneous in nature,[3] i.e., they consist of multiple components and often combine continuous and discontinuous elements. In order to study local SAR features in analog series, a data structure termed combinatorial analog graph (CAG)[5] has been developed that systematically divides analogs into subsets of compounds that only differ at defined substitution sites and, in addition, incorporates the SARI scoring scheme to identify substitution patterns that are responsible for SAR discontinuity.

SAR analysis functions and CAGs exclusively utilize similarity and potency information of active compounds as input, i.e., information encoded at the ligand level. However, SAR information contained in analog series is naturally to a

* Corresponding author. Telephone: +49-228-2699-306. Fax: +49-228-2699-341. E-mail: bajorath@bit.uni-bonn.de.
† Department of Life Science Informatics.
‡ Pharmazeutisches Institut.
§ These authors contributed equally to this work.

Table 1. Data Sets

target	PDB code	PDB ligand id	no. compds	potency range
carbonic anhydrase II	2HOC	1CN	6	0.3−9 nM
tie-2 Kinase	2P4I	MR9	8	1−399 nM
factor Xa	2BMG	I1H	20	13 nM−1 μM
factor Xa	2G00	4QC	13	0.18−88 nM
thrombin	1SL3	170	13	0.0014−940 nM

Analog series including a ligand with known X-ray structure were collected from the BindingDB[8] or the literature.[9] The "PDB code" is the PDB code of the crystallographic enzyme−inhibitor complex, the "PDB ligand id" is the unique PDB identifier of the X-ray ligand, and "no. compds" denotes the number of compounds.

large extent determined by underlying receptor−ligand interactions,[1] which are the focal point of structure-based ligand design efforts.[6] Given this close link between receptor−ligand interactions and compound activity/potency, one might perhaps expect that relationships between molecular structure and biological activity observed at the ligand level could be easily reconciled on the basis of 3D interaction information. However, it has been demonstrated that relationships between 2D similarity, 3D (binding mode) similarity, and potency of active compounds are often highly complex and difficult to predict.[7] This emphasizes the fact that interactions seen in receptor−ligand complexes are only one of several factors that determine or influence SARs, in addition to, for example, entropic effects or desolvation energy. Nevertheless, the presence of strong SAR discontinuity is often thought to be a direct consequence of compromised receptor−ligand interactions.

In order to further improve our understanding of SAR discontinuity, we have carried out an analysis of analog sets for which 3D structural information was available. This made it possible to interpret SAR information extracted from active compounds projected into chemical reference spaces at the target structural level. To these ends, series of analogs directed against different targets were systematically analyzed in CAGs in order to identify substitution patterns that were directly responsible for SAR discontinuity within these series. Key substitution sites were then analyzed on the basis of crystallographic data to map activity cliffs and rationalize SAR discontinuity within the framework of specific receptor−ligand interactions. The analysis shows that many discontinuous SAR features extracted from CAGs can be directly associated with experimental receptor−ligand interactions. However, this is not always possible, and some substitution site patterns that introduce significant SAR discontinuity in analog series cannot be explained in structural terms.

MATERIALS AND METHODS

Analog Series and X-ray Structures. Active compounds and potency data were taken from BindingDB,[8] with the exception of the thrombin inhibitor series that was taken from the literature.[9] We selected five series of structurally analogous inhibitors of four target enzymes for which one of the analogs was available in a complex X-ray structure. Analog series included inhibitors of carbonic anhydrase II (PDB code 2HOC), tie-2 kinase (2P4I), factor Xa (2BMG and 2G00), and thrombin (1SL3). Table 1 summarizes these compound data sets. Enzyme−inhibitor interactions in X-ray structures were analyzed with MOE.[10] For the structural

correlation analysis presented herein, SAR data must be obtained from enzyme-based inhibition assays, not other assay formats.

R-Group Decomposition. For each analog series, the maximum common subgraph (MCS) shared by all analogs in the series was determined. The MCS was then utilized as the invariant core structure for R-group decomposition to determine (and consistently number) corresponding substitution sites in analogs and assign sets of substituents (functional groups) to these sites. For the analog series studied herein, all substitution sites were unambiguously assigned. MCS calculation and R-group decomposition were performed using Pipeline Pilot.[11] Table 2 reports core structures, substitution sites, and R-groups for all five analog series.

Organization and Analysis of Analog Series. In order to quantify contributions of substitution sites to SAR discontinuity, we systematically divided analog series into subsets of compounds that only differed at a specific substitution site or a combination of up to three sites. For the resulting compound subsets, the SARI discontinuity score[4] was calculated, and CAGs[5] were used to organize analog subsets and visualize contributions of substitution patterns to SAR discontinuity.

SARI Discontinuity Score. SAR discontinuity within an analog series results from small chemical modifications that cause significant changes in potency. The SARI discontinuity score calculates pairwise potency differences between analogs and averages them to obtain a measure of the magnitude of potency differences among similar compounds. The pairwise potency differences are scaled by the similarity value of the corresponding compound pair in order to emphasize potency differences between highly similar compound pairs:
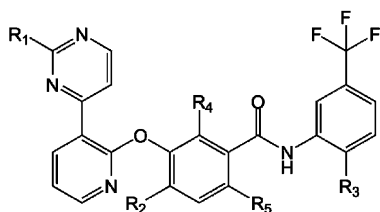
$$\text{disc} = \underset{\{(i,j)|i \neq j\}}{\text{mean}} (|P_i - P_j| \times \text{sim}(i,j))$$

Here, P_i and P_j denote the potency values of compounds i and j and $\text{sim}(i,j)$ denotes their similarity, calculated as the Tanimoto coefficient (Tc)[12] for MACCS fingerprint representations.[13] SARI scoring has been found to be rather stable for compound reference sets of varying size and different molecular representations including structural keys and topological fingerprints.[1] For analyzing analog series, the application of a similarity threshold value for calculating the discontinuity score is not required because analogs have by definition highly similar structures. Therefore, the chosen molecular representation is also not critical in this case as long as it correctly counts for the high similarity of the compared compounds (which is certainly the case for structural keys).
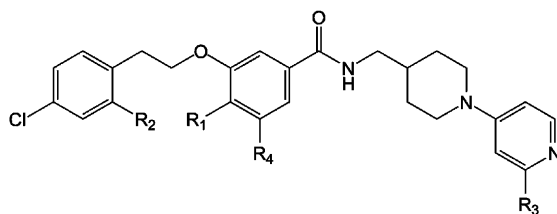
The discontinuity score is calculated for an entire analog series and all of its subsets. Scores are normalized and mapped to the value range [0,1], as described previously.[4] Score values close to 1 account for a strongly discontinuous SAR, whereas low values close to 0 reflect a low degree of SAR discontinuity. Here, scores of all compound subsets within an analog series serve as reference for normalization of the series. Thus, score distributions are characteristic of a given analog series and account for individual potency distributions.

Table 2. SAR Data for Analog Series Discussed in the Text

(a) Carbonic anhydrase II inhibitors
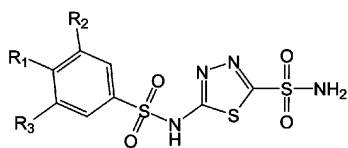
BindingDB monomer id	Potency [nM]	R1	R2	R3
10870	2	z–NH2		
10886	9			
11621	0.8	z–NH2	z–F	
11622	0.6	z–NH2		z–Cl
11625 (1CN)	0.3	z–NH2	z–F	z–Cl
11628	0.5	z–NH2	z–Cl	z–Cl
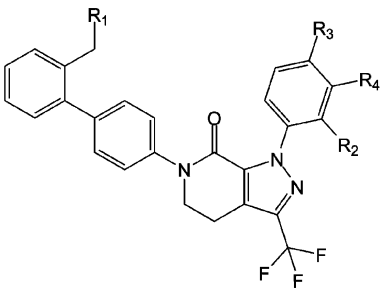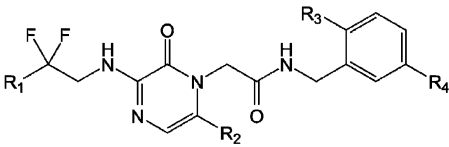
(b) Tie-2 kinase inhibitors

BindingDB monomer id	Potency [nM]	R1	R2	R3
14948	1	z–NH–	z–	
14977	153		z–	
14982	399		z–	Z–N(morpholine)
14983 (MR9)	10	z–NH–	z–	Z–N(morpholine)
14989	99	z–NH–		Z–N(piperidine)
14992	39	z–NH–	z–	Z–N(piperidine)
14993	388	z–NH–		Z–N(piperidine)
14995	4	z–NH–		N–(CH2)3–N(Z)(CH3)

(c) Factor Xa inhibitors (series 1)

BindingDB monomer id	Potency [nM]	R1	R2	R3	R4
13616	50	z–F	z–Cl		
13639	106		z–Cl		
13641	156	z–I	z–Cl		
13642	37	z–	z–Cl		
13645	29	z–NH2	z–Cl		
13646	48	z–NH–	z–Cl		
13650	588	z–C(O)NH2	z–Cl		
13651	51		z–Cl		z–C(O)NH2
13653	61	z–O–	z–Cl		
13654	233	z–O–iPr	z–Cl		
13655	125		z–Cl		z–O–
13656	50	z–O–	z–Cl		z–O–
13657	263	z–O–	z–Cl		z–O–
13661	1053	z–O–	z–Cl		Z–Br
13663	13	z–	z–Cl		z–O–
13664 (I1H)	18	z–O–	z–Cl		
13665	41	z–O–			
13667	41	z–Cl	z–Cl		
13673	57	z–O–	z–Cl	z–	
13678	950	z–O–	z–Cl		Z–OH

Table 2. Continued

(d) Factor Xa inhibitors (series 2)

BindingDB monomer id	Potency [nM]	R1	R2	R3	R4
12730	0.82	Z—N	Z—S(=O)(=O)—NH2		
12731	2.7	Z—N	Z—C(=O)NH2		
12732	12	Z—N			Z—C≡N
12733 (4QC)	0.18	Z—N			Z—C(=O)NH2
12734	20	Z—N			Z—N+(=O)O−
12735	2	Z—N			Z—NH2
12736	88	Z—N			Z—NH—C(=O)CH3
12737	54	Z—N			Z—NH—C(=O)NH2
12738	47	Z—N			Z—NH—S(=O)(=O)CH3
12739	2.6	Z—N	Z—CH2CH3		
12740	0.35	Z—N	Z—O—CH3		
12742	0.72	Z—pyrrolidine-OH			Z—C(=O)NH2
12743	0.18	Z—pyrrolidine-OH	Z—O—CH3		

(e) Thrombin inhibitors

Compound	Potency [nM]	R1	R2	R3	R4
4	12	Z—pyridyl	Z—Cl		
5	0.44	Z—pyridyl	Z—Cl		Z—Cl
13	0.45	Z—pyridyl	Z—Cl	Z—triazolyl	
14	0.01	Z—pyridyl	Z—Cl	Z—triazolyl	Z—Cl
16	0.1	Z—pyridyl	Z—Cl	Z—tetrazolyl	
17	0.0015	Z—pyridyl	Z—Cl	Z—tetrazolyl	Z—Cl
19	940	Z—pyridyl	Z—Cl	Z—tetrazolyl(NH)	
24	16	Z—pyridyl		Z—triazolyl	
25	0.24	Z—pyridyl		Z—triazolyl	Z—Cl
33	0.05	Z—pyridyl N-oxide	Z—Cl	Z—tetrazolyl	
34 (170)	0.0014	Z—pyridyl N-oxide	Z—Cl	Z—tetrazolyl	Z—Cl
35	2.7	Z—pyridyl		Z—tetrazolyl	
36	0.033	Z—pyridyl		Z—tetrazolyl	Z—Cl

For each analog series discussed in the text, molecular frameworks and consistently numbered R-groups are presented. For individual compounds, substituents and potency values (K_i or IC_{50}) are reported. Attachment atoms are labeled with "Z". Compounds from the BindingDB are identified by their unique BindingDB monomer id. Thrombin inhibitors are identified by their indices from the original publication.[9] For X-ray ligands, their PDB ligand identifier is given in parentheses.

RESULTS

Combinatorial Analog Graph Data Structure. The CAG representation organizes analog series as compound subsets having modifications exclusively at defined sites or site combinations and accounts for SAR discontinuity within subsets that can be directly attributed to modifications at the given substitution sites. Figure 1a shows a simple CAG

representation for a small set of six carbonic anhydrase II inhibitors. The compounds are also shown (in addition to Table 2), and the subsets they form at different nodes are reported, illustrating that analogs usually participate in different subsets, given the distribution of substituents. As exemplified in Figure 1a, a node might correspond to several subsets that consist of compounds that differ only at the given
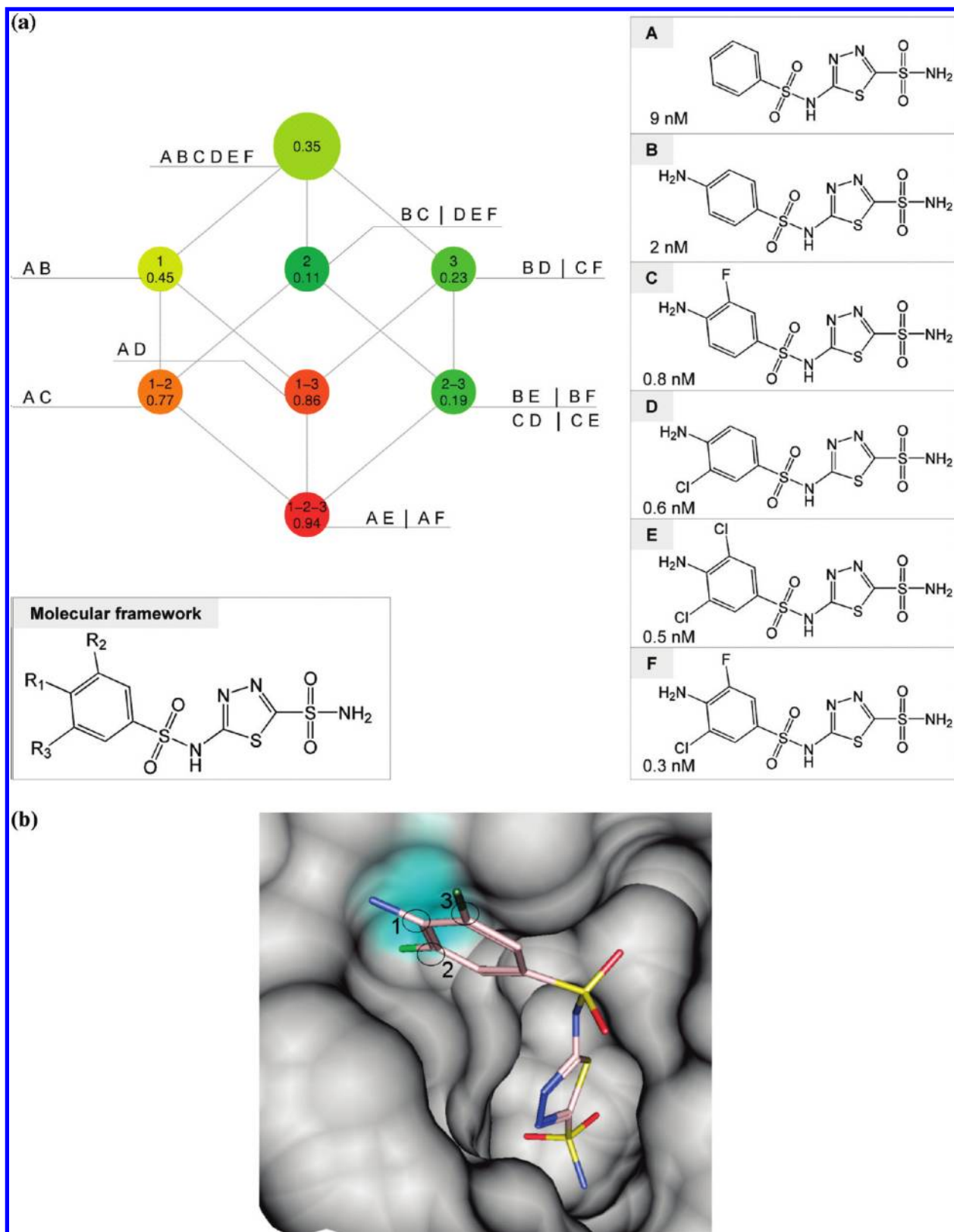
Figure 1. SAR discontinuity analysis: carbonic anhydrase II inhibitors. (a) Combinatorial analog graph for the inhibitor series. Nodes in the CAG are labeled with substitution sites and site combinations and color coded according to the degree of SAR discontinuity within the corresponding compound subsets. In this case, all inhibitor structures are shown, and the nodes also have inhibitor labels in order to illustrate the composition of overlapping compound subsets corresponding to the nodes. If more than one subset is available for a node (e.g., at nodes 2, 3, 2−3, and 1−2−3), then individual subsets are separated by a vertical line. (b) Complex crystal structure of carbonic anhydrase II with inhibitor 11 625 (numbered according to Table 2; PDB code 2HOC). In this and the following structural representations, the active site of the enzyme is depicted in solid surface representation. Selected hydrophobic pockets are shown with cyan surface coloring, and charged/polar pockets are shown with red coloring. Inhibitors are shown in stick representation using the following atom color code: light-rose, carbon; red, oxygen; blue, nitrogen; yellow, sulfur; and green, halogen. Substitution sites in inhibitors are circled and labeled.

sites but are distinguished at another site. Discontinuity scores for these subsets are calculated independently and averaged. In the CAG, the top (root) node represents the entire analog series, and each nonroot node represents a subset of compounds with different substitutions at the specified sites. Node labels identify the substitution sites and report discontinuity scores for the compound subset representing each site combination. For example, "1" and "1–2" means that compound subsets only differ at site 1 or sites 1 and 2, respectively, but are otherwise identical. Nodes are arranged in layers according to the number of substitution sites that are considered and color coded according to discontinuity scores using a spectrum from green (score 0, i.e., no SAR discontinuity) over yellow to red (score 1, i.e., maximal discontinuity). Edges are drawn from a node to all other nodes in the next layer whose substitution site combination includes all of the sites represented by the originating node. Substitution site combinations for which no compounds are available (i.e., nonexplored combinations) are shown as small white nodes. Combinations of up to three sites are systematically accounted for, and the complexity of a CAG increases with the number of individual sites. For example, for three substitution sites, one bottom node with a three-site combination is obtained (Figure 1a), but for four sites, there are four bottom nodes (Figure 3a).

Patterns of SAR Discontinuity and Mapping of Activity Cliffs. For each of the five analog series, CAG representations were generated, and in each case, substitution patterns were identified that introduced SAR discontinuity. We found that discontinuity patterns substantially varied among these analog series. In the following, the CAG representations of the different series are discussed, and substitution sites that introduce SAR discontinuity are evaluated in the context of enzyme–inhibitor interactions.

Carbonic Anhydrase II. The six analogous carbonic anhydrase inhibitors, including the X-ray ligand, range in potency from 0.3–9 nM and differ at three substitution sites. The CAG in Figure 1a clearly shows that individual modifications at single substitution sites do not lead to significant potency differences (green nodes 1, 2, and 3 at the first level). By contrast, combinations of modifications at sites 1–2 and 1–3 (but not 2–3) result in potency differences of more than 1 order of magnitude, and largest SAR discontinuity is observed for simultaneous modifications at sites 1–2–3 (red node at the bottom). Thus, given the SAR information that is extracted from overlapping compound subsets, this series would be suggested to contain substitutions that depend on each other and act in concert.

The X-ray structure of the enzyme–inhibitor complex in Figure 1b shows that substitutions at site 1 point outside the active site and are, thus, not expected to interact with the protein. Furthermore, the complex reveals information that could not be deduced from SAR analysis of this analog series. Importantly, for substitutions at sites 2 and 3 at the phenyl ring that is freely rotatable, only one small hydrophobic binding pocket exists formed by residues Val135, Leu198, Pro202, and Leu204. Filling this pocket represents an activity cliff for strong inhibition. However, this can be accomplished by a halogen substituent at either site 2 or 3. Thus, substitutions at these sites do not act in a concerted manner, as suggested by analyzing the series, but in an alternative way. This information could not be deduced from

the CAG because halogen-substituted compounds at site 2, 3, or both sites have comparable potency, consistent with the structural picture, and influence SAR discontinuity in similar ways.

Tie-2 Kinase. This series consists of eight analogs that are ATP site-directed inhibitors. They differ at five substitution sites and fall into the potency range of 1–399 nM. As illustrated by the CAG in Figure 2a, in this series, many potential combinations of substitution sites are currently unexplored (shown as "empty" nodes). However, substitution site 1 emerges as a prominent hotspot for individual modifications. For example, the addition of a methylamine at site 1 increases potency by up to 2 orders of magnitude (e.g. compounds 14948 and 14977, Table 2). This can be easily explained considering the structure and inhibitor complex shown in Figure 2b. This site 1 substituent forms a strong hydrogen bond to the backbone carbonyl oxygen of Ala905, at the bottom of the pocket surrounding site 1. The interaction with this residue that is conserved in many kinases represents a prominent activity cliff for ATP site-directed inhibition.

Furthermore, modifications at sites 2–3 and 2–4 introduce moderate SAR discontinuity, but combinations of the same modifications at sites 2–3–4 yield a high degree of discontinuity. Thus, modifications at sites 2–3 and 2–4 have additive effects. Modifications of sites 2 and 4 include the presence or absence of a methyl group and only have a minor SAR effect. However, in compounds corresponding to node 2–4, the methyl position is exchanged between sites 2 and 4, which results in moderate SAR discontinuity. This is the case because the hydrophobic pocket facing site 4 is smaller than the one facing site 2 (Figure 2b), and the methyl group at site 4 is likely to form an unfavorably close contact with the Phe296 side chain, which reduces compound potency by an order of magnitude.

Moreover, in compounds corresponding to node 2–3–4, the methyl position is also switched between sites 2 and 4, and site 3 contains a piperazine or morpholine group or no substituent. These modifications at site 3 are also present in compounds corresponding to node 2–3, but only lead to moderate SAR discontinuity. However, simultaneous modifications at sites 2, 3, and 4 act in concert and lead to a considerable degree of discontinuity. Nodes 3–5 and 3–4–5 also display moderate and high discontinuity, respectively, which primarily results from the change of a cyclic to an acyclic substituent at site 3 (Table 2). Preferences for substituents at site 3 are not apparent from the structure, except that crystallographic temperature factors indicate significant protein backbone flexibility in the region, which might give rise to induced fit effects.

Factor Xa, Series 1. This series contains 20 analogs that differ at four substitution sites and span a large potency range of 13 nM–1 μM. In this case, substitution site 3 forms a prominent SAR hotspot (Figure 3a). SAR discontinuity at this site is largely determined by the presence or absence of a hydroxyl substituent at the pyridine moiety that is engaged in strong π–π stacking interactions with aromatic binding site residues (Figure 3b). Addition of the OH-group to the pyridine enables tautomerization, which destabilizes its π electron system and diminishes the stacking interaction. A weakly potent analog having a bromine substituent at site 4 also introduces moderate SAR discontinuity in this compound
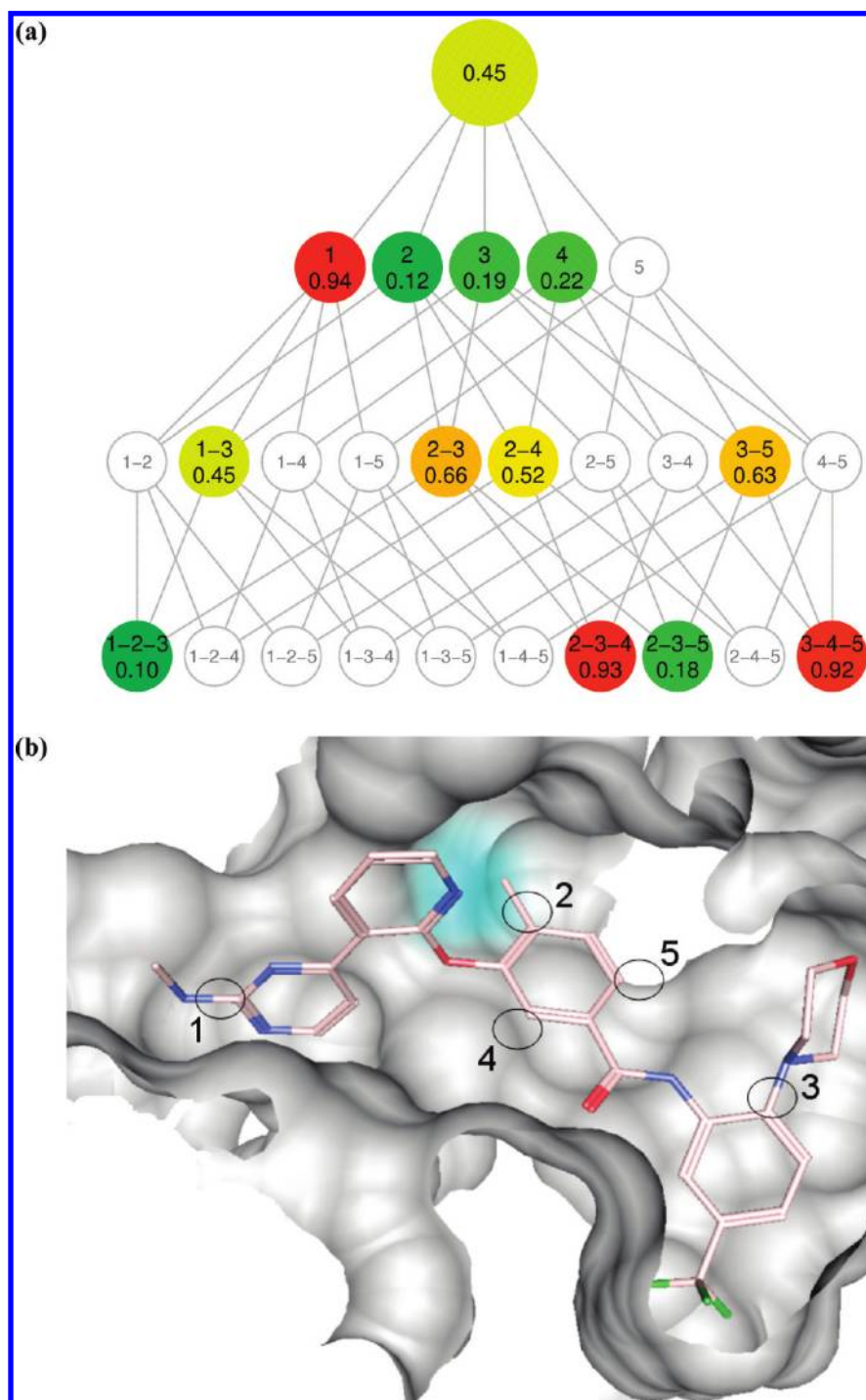
ANALYSIS OF STRUCTURE−ACTIVITY RELATIONSHIPS

J. Chem. Inf. Model., Vol. 49, No. 10, 2009 **2185**

Figure 2. Tie-2 kinase inhibitors. (a) CAG for a set of eight inhibitors of tie-2 kinase with substitutions at up to five different sites. (b) X-ray crystal structure of tie-2 kinase in complex with inhibitor 14 983 (numbered according to Table 2; PDB code 2P4I).

subset because a bromine at site 4 is too large at the entrance of the hydrophobic pocket harboring site 1 substituents (Figure 3b) and imposes steric constraints. Hence, nodes including substitution sites 3 and 4 make the overall largest contributions to SAR discontinuity.

Factor Xa, Series 2. This alternative inhibitor series consists of 13 highly potent analogs (0.18−88 nM) that differ at four substitution sites. Their CAG, presented in Figure 4a, mirrors SAR features that are significantly different from the factor Xa inhibitor series discussed above. Series 2 displays a well-defined pattern of compound subsets representing discontinuous SARs. All of these compound subsets involve substitutions at site 4 that result in potency differ-

ences of up to 2 orders of magnitude. Figure 4b shows the X-ray structure of the complex containing the inhibitor that has the most favorable substituent at this position. The structural representation confirms that site 4 substituents reach into the S1 pocket in the active site of factor Xa that contains residue Asp189 at the bottom. Interactions with this residue (or corresponding residues) represent a critical activity cliff and are a hallmark of trypsin-like serine protease inhibition. The inhibitor shown in Figure 4b fills this pocket and forms a hydrogen-bonding network with residues Asp198 and Gly218, which is consistent with its high potency. Compared to this interaction constraint, substitutions at other
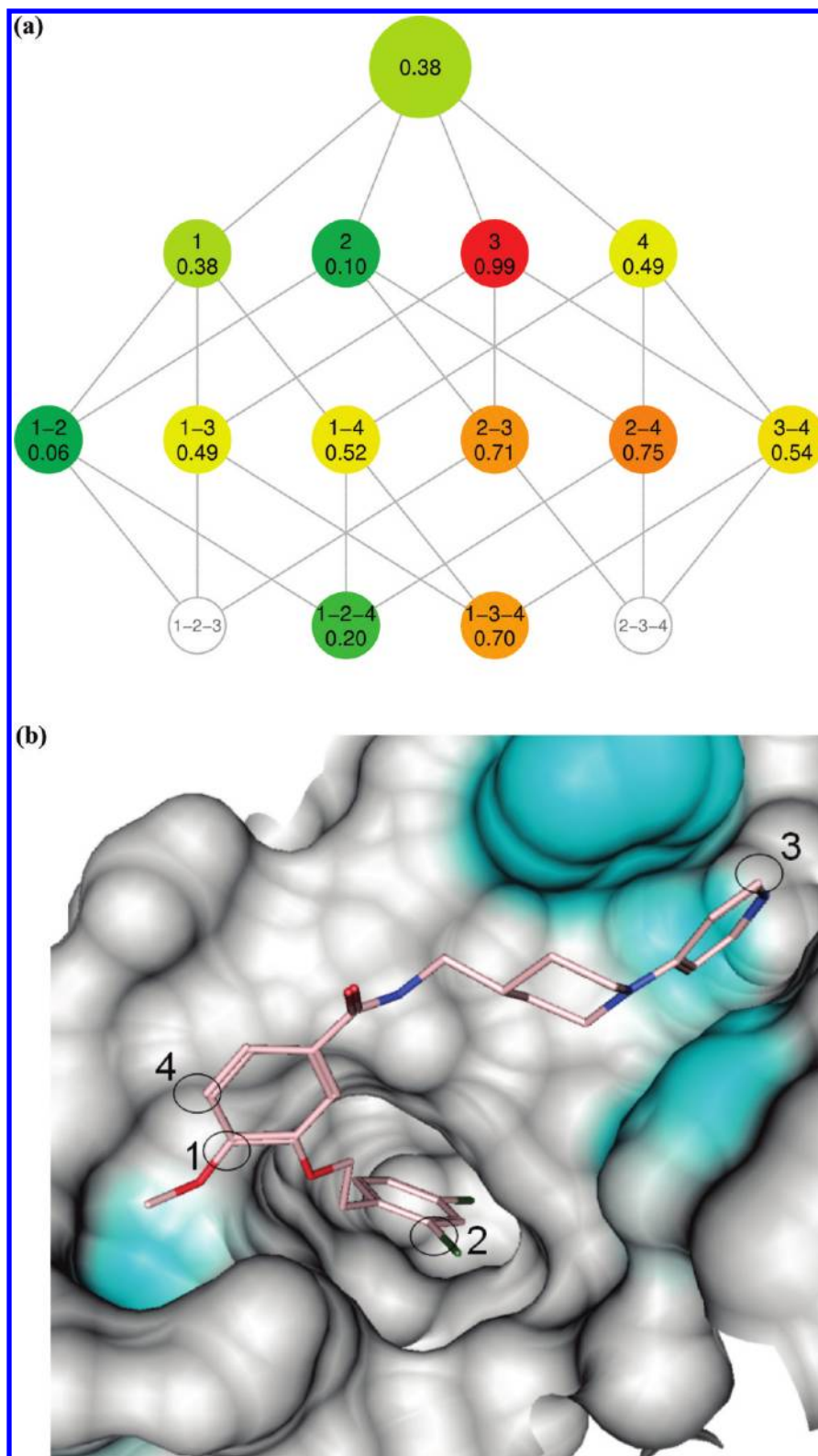
Figure 3. Factor Xa inhibitor series 1. (a) CAG for a set of 20 analogous factor Xa inhibitors with four substitution sites. (b) Crystal structure of inhibitor 13_664 bound to factor Xa (PDB code 2BMG).

sites or site combinations introduce considerably less SAR discontinuity.

Thrombin. This set of thrombin inhibitors contains 13 analogs having four substitution sites and potency in the range of 0.0014−940 nM. In this case, most significant SAR discontinuity is observed for combinations of modifications involving sites 2, 3, and 4 (Figure 5a). Among individual sites, only substitutions at site 4 lead

to notable discontinuity, which is due to an analog having a chlorine substituent at site 4 (in the absence of a chlorine at site 2). The simultaneous addition of chlorine substituents at both sites (node 2−4) further increases SAR discontinuity by causing a potency increase of several orders of magnitude. The strong discontinuity at node 3−4 is mainly due to variations at site 3 (triazole, tetrazole, or no substituent). Analogs at nodes 1−2−3, 1−3−4, and
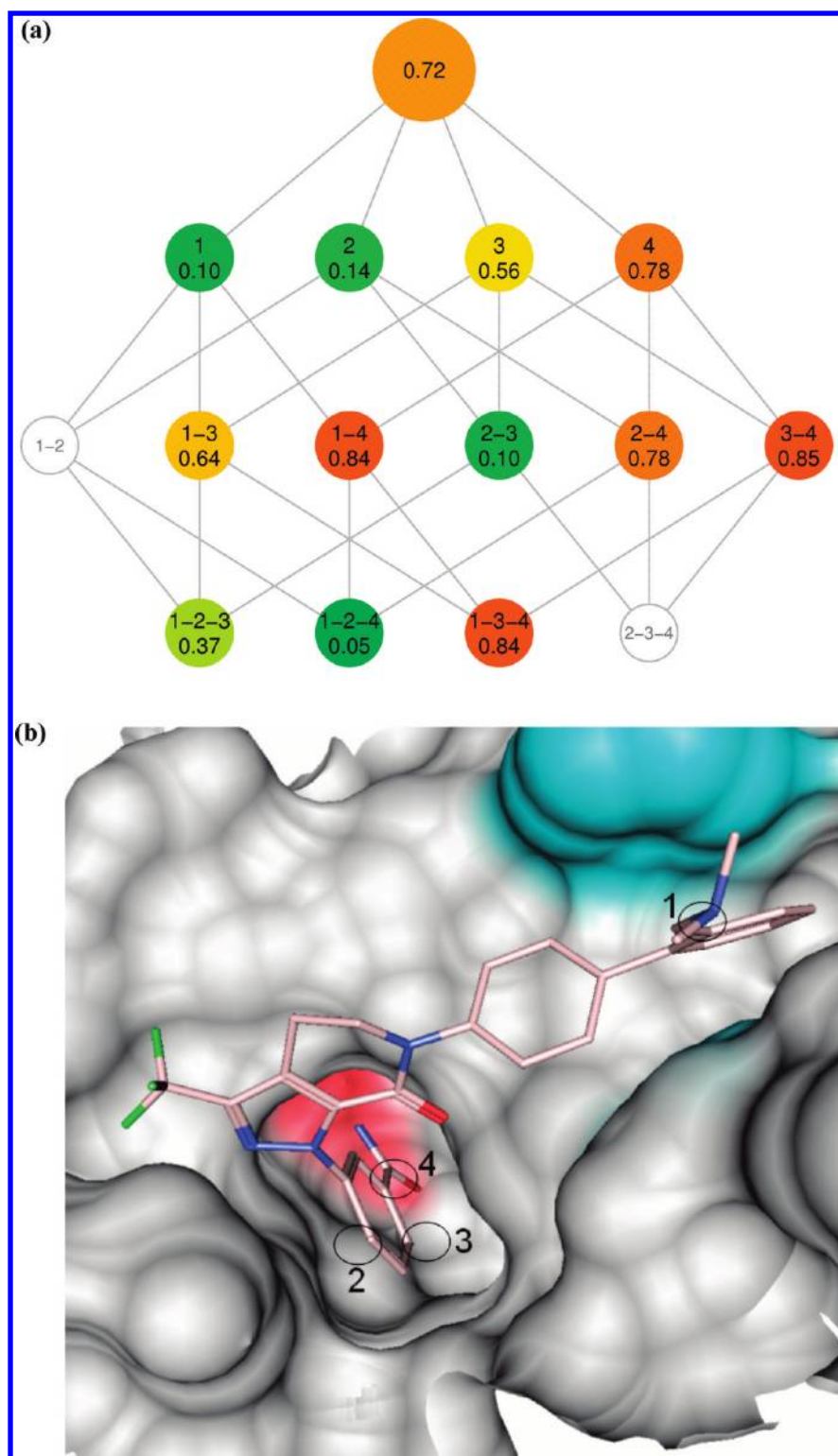
Figure 4. Factor Xa inhibitor series 2. (a) CAG for another analog series consisting of 13 factor Xa inhibitors with four substitution sites. (b) Complex crystal structure of factor Xa with inhibitor 12 733 (PDB code 2G00).

2−3−4 combine the modifications described above that strongly affect potency.

The X-ray structure in Figure 5b contains the most potent analog having chlorine substituents at both sites 2 and 4. Substituents at site 3 are partly solvent exposed, and the significant SAR discontinuity introduced by site 3 variations reflected in the CAG is difficult to explain in terms of interactions seen in the X-ray structure. Site 4 points into the S1 pocket, and interactions involving key residues

within the S1 pocket represent a pronounced activity cliff. In this analog series, however, the S1 pocket is not occupied with a positively charged group but with a chlorine site 4 substituent that strongly interacts with the π system of Tyr228. The chlorine substituent at site 2 also fills a small hydrophobic pocket, and the simultaneous presence of both chlorine substituents leads to a highly potent analog and a strong SAR discontinuity detected in the CAG. The cooperative nature of these two sites is
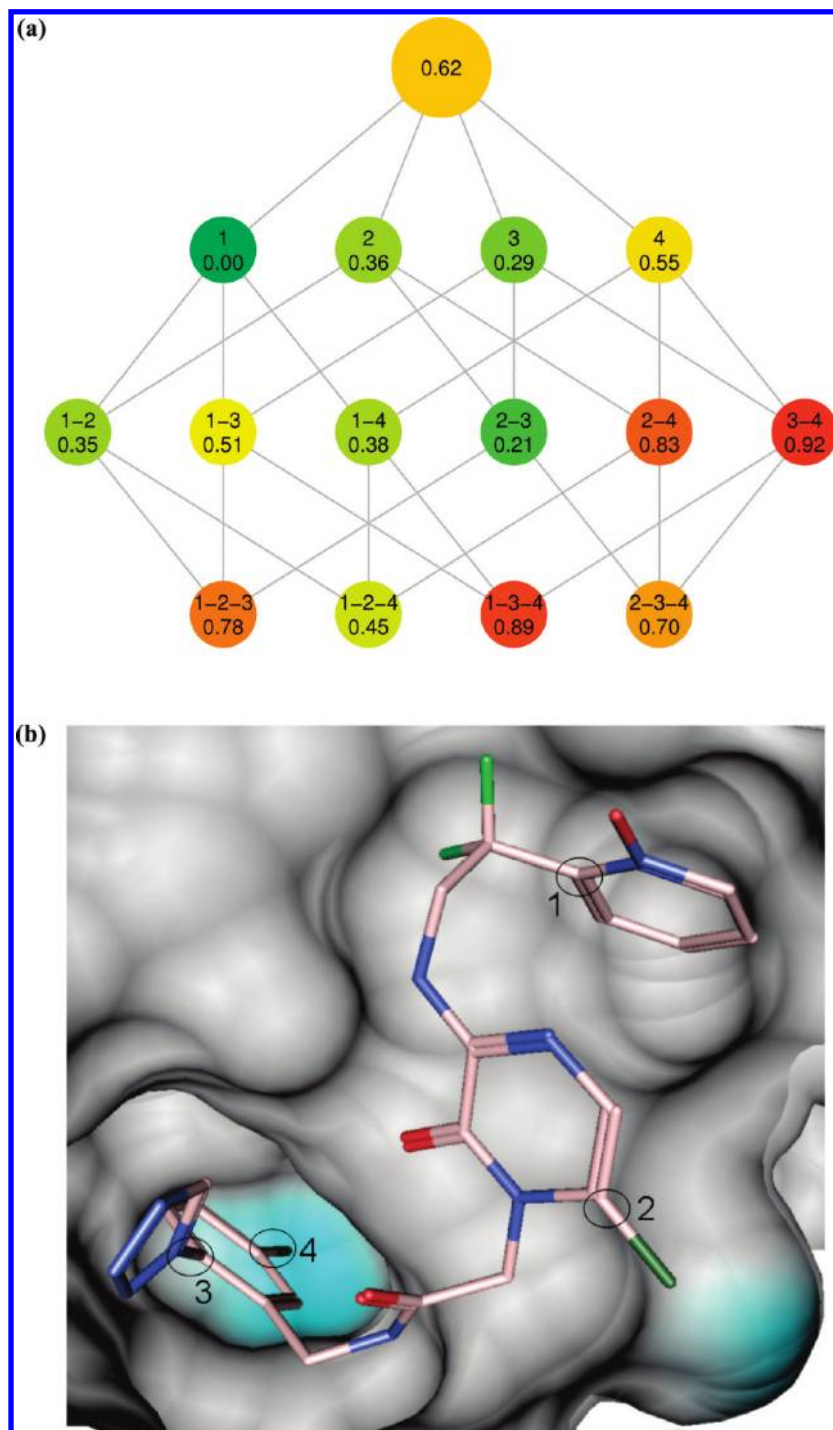
Figure 5. Thrombin inhibitors. (a) CAG for 13 analogous thrombin inhibitors with four substitution sites. (b) Crystal structure of inhibitor 34 bound to the active site of thrombin (PDB code 1SL3).

again hard to reconcile on the basis of the structure. Thus, taken together, the strong SAR discontinuity introduced by 2−4 and 3−4 substitutions in this series of thrombin inhibitors, as revealed by CAG analysis, could not have been deduced from interaction patterns in the X-ray structure.

DISCUSSION

The primary motivation of our study has been to analyze SAR information contained in analog series and interpret the results at the level of target−ligand interactions seen in complex X-ray structures. We hoped to better understand

how SAR discontinuity detected at the ligand level is reflected by interaction information derived from complex crystal structures and, if possible, arrive at a structural interpretation of individual activity cliffs. The systematic extraction of SAR information from compound series is a nontrivial task. For this purpose, we have developed combinatorial analog graphs that, on the basis of R-group decomposition and SARI scoring, make it possible to organize compound sets and identify substitution patterns that are responsible for activity cliffs and induce SAR discontinuity. These graph representations have been applied here to analyze selected analog series directed against

ANALYSIS OF STRUCTURE–ACTIVITY RELATIONSHIPS

J. Chem. Inf. Model., Vol. 49, No. 10, 2009 **2189**

different targets. Care has been taken to select series for which X-ray structural information was available, and that shared large maximum common subgraphs and multiple substitution sites. This explains why the analog series studied here were of relatively small size because not very many analogs could be identified that met these requirements. However, the selected series were well-suited to study SAR discontinuity resulting from minor variations of substitutions in otherwise identical molecules. For each of these series, compound subsets with well-defined substitution patterns were identified that introduced significant degrees of SAR discontinuity, and in a number of cases, activity cliffs could be readily mapped in X-ray structures. However, we have also found that discontinuity patterns substantially differed between analog series, as exemplified by the two factor Xa inhibitor series, and that it was not possible in all cases to rationalize SAR determinants at the structural level. Thus, although often assumed, there is not always a consistent and close correspondence between SAR discontinuity and compromised receptor–ligand interactions. The relationship between SAR information encoded in analog series and receptor–ligand interactions seen in receptor–ligand structures is more complex. In some instances, individual SAR hotspots revealed in graph representations could be easily associated with critical interactions as, for example, in the case of the S1 site in thrombin or the tautomerization effects in factor Xa inhibitors. By contrast, the SAR features of, for example, substitution site combinations in thrombin inhibitors could not be rationalized in structural terms. On the other hand, structural analysis helped to clarify SAR ambiguities detected at the ligand level as in the case of carbonic anhydrase II inhibitors. Thus, taken together, the results of our study also point at the complementary nature of ligand-based SAR and structural analyses.

CONCLUSIONS

In our study, SAR discontinuity information deduced from combinatorial analog graphs of different inhibitor series has been analyzed in light of receptor–ligand interactions in complex crystal structures, which has made it possible to explore activity cliffs at the small and macromolecular level. Although many effects of substitutions at defined sites in inhibitors could be rationalized in structural terms, SAR discontinuity detected in analog series could not only be attributed to the presence or absence of specific receptor–ligand interactions. However, structural interpretation helped to better understand the origin of SAR discontinuity in cases where ligand-based analysis was insufficient. Clearly, information provided by the systematic comparison of analogs and the analysis of complex crystal structures was highly complementary in a number of cases. Approaches for the extraction of SAR information from compound data sets should provide attractive starting points for the detection of activity cliffs and further exploration of local SAR patterns and discontinuity at the target structural level.

REFERENCES AND NOTES

(1) Peltason, L.; Bajorath, J. Systematic Computational Analysis of Structure-Activity Relationships: Concepts, Challenges and Recent Advances. *Future Med. Chem.* **2009**, *1*, 451–466.
(2) Maggiora, G. M. On Outliers and Activity Cliffs - Why QSAR Often Disappoints. *J. Chem. Inf. Model.* **2006**, *46*, 1535.
(3) Bajorath, J.; Peltason, L.; Wawer, M.; Guha, R.; Lajiness, M. S.; Van Drie, J. H. Navigating Structure-Activity Landscapes. *Drug Discov. Today* **2009**, *14*, 698–705.
(4) Peltason, L.; Bajorath, J. SAR Index: Quantifying the Nature of Structure-Activity Relationships. *J. Med. Chem.* **2007**, *50*, 5571–5578.
(5) Peltason, L.; Weskamp, N.; Teckentrup, A.; Bajorath, J. Exploration of Structure-Activity Relationship Determinants in Analogue Series. *J. Med. Chem.* **2009**, *52*, 3212–3224.
(6) Jorgensen, W. L. The Many Roles of Computation in Drug Discovery. *Science* **2004**, *303*, 1813–1818.
(7) Peltason, L.; Bajorath, J. Molecular Similarity Analysis Uncovers Heterogeneous Structure-Activity Relationships and Variable Activity Landscapes. *Chem. Biol.* **2007**, *14*, 489–497.
(8) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: A Web-Accessible Database of Experimentally Determined Protein-Ligand Binding Affinities. *Nucleic Acids Res.* **2007**, *35*, D198–201.
(9) Young, M. B.; Barrow, J. C.; Glass, K. L.; Lundell, G. F.; Newton, C. L.; Pellicore, J. M.; Rittle, K. E.; Selnick, H. G.; Stauffer, K. J.; Vacca, J. P.; Williams, P. D.; Bohn, D.; Clayton, F. C.; Cook, J. J.; Krueger, J. A.; Kuo, L. C.; Lewis, S. D.; Lucas, B. J.; McMasters, D. R.; Miller-Stein, C.; Pietrak, B. L.; Wallace, A. A.; White, R. B.; Wong, B.; Yan, Y.; Nantermet, P. G. Discovery and Evaluation of Potent P1 Aryl Heterocycle-Based Thrombin Inhibitors. *J. Med. Chem.* **2004**, *47*, 2995–3008.
(10) *Molecular Operating Environment (MOE)*, Version *2008.10*; Chemical Computing Group, Inc.: Montreal, QC Canada, 2008.
(11) *Scitegic Pipeline Pilot*, Student Ed., Version *6.1*; Accelrys, Inc.: San Diego, CA, 2007.
(12) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(13) *MACCS Structural Keys*; Symyx Software: San Ramon, CA.

CI900243A