

Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers?

Ki Hwan Kim

Received: 12 March 2007 / Accepted: 13 June 2007 / Published online: 24 July 2007
© Springer Science+Business Media B.V. 2007

Abstract Structure-activity relationship (SAR) and/or quantitative structure-activity relationship (QSAR) studies play an important role in a lead optimization of drug discovery research. When there is a lack of ligand-bound protein structural information, one of the assumptions in SAR and QSAR studies is that similar analogs bind to the same binding site in a similar binding mode. In such studies, outliers have often been observed, especially in QSAR. However, most of these studies have focused their attention on the development of QSAR and left outliers unattended. We searched ligand-bound X-ray crystal structures from the protein structure database to find evidences that could indicate a possible source of outliers in SAR or QSAR. Our results showed the possibility of conformational changes in a flexible binding site as one possible source of outliers.

Keywords SAR · QSAR · Outliers · Multiple binding mode · Flexible binding site · X-ray crystal structure · Source of outliers

Introduction

In drug discovery research, structure-activity relationship (SAR) studies play an important role in the optimization

process of a lead compound. Studies of SAR and quantitative structure-activity relationship (QSAR) assume that structurally similar analogs bind to the same binding site in a similar binding mode. Ligand-bound X-ray crystal structures have proven this true in many cases. Since the first proposal of QSAR technique over 40 years ago, more than 18,000 physical and biological QSAR equations have been reported. Hansch and his co-workers have collected them into the world's largest QSAR database called C-QSAR [1]. This database is available through the BioByte Corporation with the Bio-Loom interface [2].

Outliers are often observed in SAR or QSAR. We noted that there is a significant number of cases where one or more compounds were left out to develop the reported QSARs in the C-QSAR database [3]. Outliers of QSARs can be very important and interesting, especially when the observed biological activity is higher than the predicted one by the QSAR.

There could be various reasons for the observed outliers. Outliers may be present due to the inappropriate calculation of the parameter values used. There may be a lack of certain descriptors or parameters to describe the QSAR for the entire compounds. The mathematical model may not be appropriate. A different mode of mechanism may even be a reason.

In the previous study, [3] we focused our initial efforts on finding crystal structural evidences that might indicate a possible source of outliers in SAR or QSAR. Based on the observed ligand-bound X-ray crystal structures of various proteins, we suggested that unusual binding modes of structurally similar analogs could be a possible source of outliers. In this study, we focus our attention to another possible source of outliers in SAR and QSAR—flexible binding sites.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-007-9126-y) contains supplementary material, which is available to authorized users.

K. H. Kim (✉)
Hope Drug Discovery Research Laboratory,
260 Southgate Drive, Vernon Hills, IL 60061, USA
e-mail: pkhkim@gmail.com

Experimental section

C-QSAR database [1] searching

The database was searched with a query ‘elastase inhibitor’, ‘rhinovirus inhibitor’, or ‘nitric oxide synthetase inhibitor’ using the BioByte Corporation’s Bio-Loom interface [2]. The resulting list of the search was further examined for individual QSAR summary for the number of outliers, the number of outliers with the deviation greater than 3 times its standard deviation, and number of outliers with the deviation greater than 1.0 or 2.0 positive and negative deviations. The percentage of the outliers with respect to the total number of compounds in each dataset and the average percentage of the outliers from the entire datasets were then calculated.

RCSB protein data bank [4] searching

The database was searched for most of the compounds presented in this paper based on the literature information concerning flexible binding sites. The resulting pdb files from the searches were then further examined for their flexible binding sites as well as binding modes by superimposing the protein structures in each sub-group. The multiple sequence alignment for protein structure comparison was done using the ClustalW program of the EMBL-EBI [5].

Molecular graphics

All the figures were generated using the UCSF Chimera molecular modeling program (beta version 1) [6] using the multiple sequence alignments obtained from the ClustalW described above or structure comparison tool of Chimera.

Results and discussion

Outliers in QSAR equations

In order to examine the frequency of outliers in QSAR, we searched the C-QSAR database for all the reported QSAR equations using a particular search query. From a search query ‘carbonic anhydrase’ as a test case, 60 QSAR equations were previously found [3]. It was interesting to learn that only six out of these 60 QSAR equations utilized all compounds available for the QSAR studies. The other 54 equations had one or more outliers. The number of equations with outliers was more than what we had initially expected. Three additional queries were used in the subsequent searches to find out whether such a high occurrence of outliers in QSAR is common: ‘elastase inhibitor’, ‘rhinovirus inhibitor’, and ‘nitric oxide synthetase inhibitor’. Forty-seven QSAR equations were obtained from these searches (Table 1 and the Supplemental Material). The biological activities for these QSARs are the inhibitory activity ($\log 1/C$), K_i inhibition of the enzyme ($\log 1/K_i$), or equilibrium constant ($\log k$). In general, the quality of the reported equations is very good, judged by their standard deviations and squared correlation coefficients. Similar to the previous results [3] (i.e., 60 QSARs for the ‘carbonic anhydrase’), a majority of the QSAR equations have one or more outliers. Only one out of the 47 QSARs was reported not to have any outlier.

For the 47 QSARs, the percentage of the outliers with respect to the total number of compounds in each of the datasets ranges from 3% up to 36%. Twenty-six of the 47 datasets have 20% or more outliers; the average percentage of the outliers from the entire 47 datasets is 20.1%. This means that on the average, two out of every 10 compounds in each dataset do not fit with the QSAR model. These results are very similar to those from the 60 datasets of

Table 1 QSAR equations searched for ‘elastase inhibitor’, ‘rhinovirus inhibitor’, or ‘nitric oxide synthetase inhibitor’ from C-QSAR database

No	Eq. no. ^a	n ^b	out ^c	% ^d	+; − ^e	+; − ^f	+; − ^g	s.d. ^h	R ²ⁱ	Opt ^j	Activity	Equation
<i>Elastase inhibitor</i>												
3	12023	9	3	25	0; 2	–	0; 1	0.118	0.953	1.686	$\log 1/C =$	$-3.57 \cdot \text{Clog}P + 3.91 \cdot \text{BILIN}(\text{Clog}P) + 5.87$
11	7860	18	6	25	4; 0	3; 0	1; 0	0.281	0.874		$\log k_2 =$	$-5.44 \cdot \text{MgVol} + 2.01 \cdot \text{CPI}_Y + 3.77 \cdot I + 12.19$
<i>Rhinovirus inhibitor</i>												
24	12174	7	4	36	4; 0	2; 0	1; 0	0.239	0.941	16.046	$\log 1/C =$	$2.98 \cdot \text{CMR} - 3.34 \cdot \text{BILIN}(\text{CMR}) - 38.69$
32	12173	8	3	27	2; 1	0; 1	2; 0	0.300	0.933	4.418	$\log 1/K_i =$	$250.36 \cdot \text{MgVol} - 28.34 \cdot \text{MgVol}^2 - 544.65$

For all 47 equations, see the supplemental material

^a Equation number in C-QSAR database; ^b Number of compounds included in the correlation; ^c Number of outliers; ^d Percentage of outliers with respect to the total number of compounds available for QSAR; ^e Number of outliers with the deviation greater than three times of s.d. Positive deviation, Negative deviation; ^f Number of outliers with the deviation greater than 1.0. Positive deviation, Negative deviation; ^g Number of outliers with the deviation greater than 2.0. Positive deviation, Negative deviation; ^h Standard deviation; ⁱ Squared correlation coefficient; ^j Optimum or minimum value of the parabolic or inversed parabolic equation

‘carbonic anhydrase’ QSARs [3]. Again, the percentage of outliers is very high. Such a high percentage of outliers strongly support the importance of outliers.

Eighty-one out of 118 outliers from the 47 QSARs have the standard deviations greater than three-times the standard deviation of the corresponding correlations (Such compounds are normally considered an outlier). The number of compounds with a positive deviation (the compound calculated to be less potent than the observed) and the number with a negative deviation (the compound calculated to be more potent than the observed) are not much different: 48 vs. 33. This indicates a somewhat skewed, but close to the ‘normal’ distribution of the outliers. Thirty-two of the 118 compounds have the standard deviations greater than one-logarithm unit (10-fold deviation), and 12 of the 118 have the standard deviations greater than two-logarithm unit (100-fold deviation). For these outliers, the number of compounds with a positive versus negative deviation is also not much different: 19 to 13, and 7 to 5, respectively.

Four of the 47 correlation equations require special comments. One compound in C-QSAR equation number 12023 (No. 3 in Table 1) has the deviation of 5.41 for the correlation derived from nine compounds. Something is unusual about this compound, the equation, or the predicted value of this compound from the equation. Three others are C-QSAR equation No. 7860 (No. 11 in Table 1), equation No. 12174 (No. 24 in Table 1), and equation No. 12173 (No. 32 in Table 1). Three or four outliers in these sets were calculated to be 10 to 100-times more/less potent than the experimentally measured activity value. Their deviations are much larger in magnitude than usual. Something may be missing or unaccounted for, or something may be happening in the binding modes or in the binding site of the protein.

Figure 1 is a graphic summary of the number of data sets with or without outliers. It also displays the number of outliers from the entire 107 QSAR data sets (60 data sets from the previous study [3] and 47 from this study) from the searches with the ‘carbonic anhydrase’, ‘elastase inhibitor’, ‘rhinovirus inhibitor’, and ‘nitric oxide synthase inhibitor’ queries. These results confirm our previous observations of the high rate of outliers in QSARs [3].

What are the possible sources of outliers in SAR or QSAR? It would be nice if one could explain the high percentages of observed outliers. In our previous study, we suggested that unusual binding modes of ligands could be a source of outliers. Additional examples for unusual binding modes of structurally similar compounds have been observed (Unpublished results). In this study, we attempted to search another possible source of outliers suggested by the crystal structure data from the RCSB protein structure databank [4]—the flexible binding sites in protein structures.

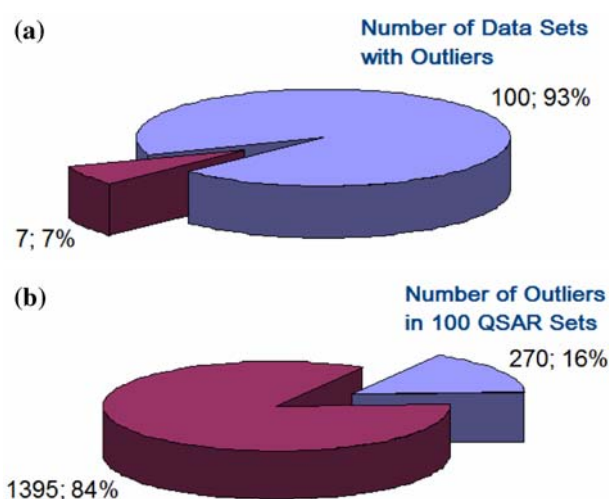


Fig. 1 (a) The number of data sets with outliers (100; 93%) and without outliers (7; 7%) in the 107 QSAR data sets (60 data sets from the previous study [2] and 47 from this study). Ninety-three percent of the 107 QSAR data sets are with outliers. (b) The number of compounds included (1395; 84%) and outliers (270; 16%) in the 107 QSAR data sets. Sixteen percent of the compounds in the 107 QSAR data sets are outliers

Cases of unusual binding modes of ligands

When ligand–protein interactions are involved, but there is a lack of ligand–protein structure information, one usually assumes that analogous molecules of interest bind to the same binding site in an essentially identical binding mode. The observed biological activity is due to the interaction of the ligand with the protein. Therefore, when the orientation of the functional group of one or more ligands is different, the descriptors for that functional group are not calculated to reflect the same ligand–protein interactions. After searching the RCSB protein databank, we found numerous examples showing how even a very closely related analog binds in a significantly different manner. Such X-ray structures may provide clues for a possible source of at least some of the outliers observed in QSARs. Many examples and detailed discussions for the cases of unusual binding modes of structurally analogous compounds were presented in our previous study [3]. All of these examples suggested that unusual binding modes of some compounds were a possible source of outliers in QSARs.

Cases of flexible binding sites as a possible source of outliers in SAR and QSAR

Proteins are dynamic molecules and often undergo conformational change with similar energies upon ligand binding. Protein flexibility is fundamental to understanding their biological effects, binding site location, binding orientation, metabolism and transport [7–13]. Protein

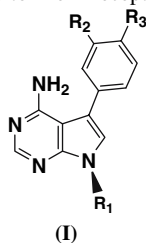
flexibility is also fundamental to understanding SAR and/or QSAR.

Protein flexibility is resulted from conformational changes in proteins. Conformational changes can be grouped into two: loop motion and domain motion [12]. In loop motion, the loops adopt different conformations upon ligand binding. Flexible loop regions often have a critical functional role in the case of enzymes. Domain motions are involved in nearly all large proteins [13]. Ligand-induced hinge motion and ligand-induced shear motion [7, 13] are the two common mechanisms for domain movement.

In the cases of unusual binding modes of ligands studied previously, [3] one or more of the ligands had a different binding mode to the respective protein. Such different binding modes of ligands may show up as outliers in SAR and/or QSAR. Even though some ligands had different binding modes, the backbone structures of the ligand-bound

proteins remained essentially identical. (There may have been some movement of one or more of the amino acid side chain at the binding site, which is involved in the ligand–protein interactions.) If the analogous ligands bind in a same binding mode to the same binding site, is an identical binding mode enough and sufficient in order to develop a correlation without expecting any outliers? Further examinations of various ligand-bound crystal structures indicate that this may not be the case. When ligand–protein interactions are involved, the observed biological activity is due to the interactions of the ligand with the protein. As in the case of unusual binding modes for structurally analogous compounds, the descriptors for the functional group of the ligands are not calculated to reflect the same ligand–protein interactions if the binding site of the protein became different for different ligands. Therefore, ligands with different shapes of binding site due to protein flexibility could

Table 2 Structures of pyrrolopyrimidinylamine analogs bound to Tie-2 receptor tyrosine kinase



Compd. no.	Structure	PDB/color in Fig. 2	Compd. no.	Structure	PDB/color in Fig. 2
1-1		Inhibitor I [14]/cyan	1-2		Inhibitor II [14]/pink
1-3		Inhibitor III [14]/gray	1-4		Inhibitor IV [14]/green

The xyz coordinates of the PDB format for these inhibitor-bound Tie-2 structures are reported without detailed resolution and R-values for the corresponding X-ray structural determination in the patent literature [14]

become outliers, even though they are analogous compounds with an identical binding mode.

Several examples illustrating this point are summarized below. They show a different shape of binding site caused by conformation changes due to a loop or domain movement. These examples suggest that changes in the protein conformation in a flexible binding site may be another source of outliers in SAR and/or QSAR. In the first two cases, the binding modes of the ligands are very similar to each other. In the next three cases in addition to the flexible binding site, the binding modes of the ligands are different.

Example 1: Pyrrolopyrimidinylamine analogs as receptor tyrosine kinase Tie-2 inhibitors

The phosphorylation reaction transfers a phosphate group from ATP to a tyrosine (in tyrosine kinases) or serine or threonine (in serine/threonine kinases) residue of the target protein substrate. Tie-2 is a member of the endothelial cell specific receptor tyrosine kinases (RTK), and one of over 400 known protein kinases (PK) that catalyze the phosphorylation of target protein substrates. There are at least nineteen distinct RTK subfamilies identified.

Bump et al. [14] studied the binding modes of Tie-2 inhibitors crystallographically. In their X-ray study, a catalytically inactive mutant of human Tie-2 was used, which included residues 802–1124 of human Tie-2, except that aspartic acid residue 964 was replaced with asparagine. Table 2 lists the X-ray crystal structures of pyrrolopyrimidinylamine analogs (structural type I) bound to Tie-2 receptor tyrosine kinase.

Figure 2 represents stereo-pair views of four pyrrolopyrimidinylamine analogs (structural type I; Compounds 1-1 in cyan, 1-2 in pink, 1-3 in gray, and 1-4 in green) bound to receptor tyrosine kinase Tie-2. In this figure and those that follow, the backbone of the protein structures are first overlapped using the sequence alignment program ClustalW of EMBL-EBI, [5] or using the UCSF Chimera molecular modeling program [6]. Only the ligands are sometimes shown for a simplicity reason. Figure 2a shows that these four compounds essentially have an identical binding mode. Fig. 2b shows that the conformation of the glycine-rich loop (P loop, nucleotide-binding loop) of Tie-2 shown on the left side is significantly different from each other, even though the superposition of these four inhibitors is very similar. Figure 2c shows a superposition of Compounds 1-3 and 1-4 bound to Tie-2, whose ligand-bound protein structures are shown in Fig. 2d. Figure 2d shows that the positions of Asp162 and Phe163 residues (of the highly conserved Asp-Phe-Gly motif of Tie-2) from the Compound 1-3-bound (gray) and 1-4-bound Tie-2 (green) are almost opposite in direction. This displays the significant changes in the binding pocket for the binding of these inhibitors.

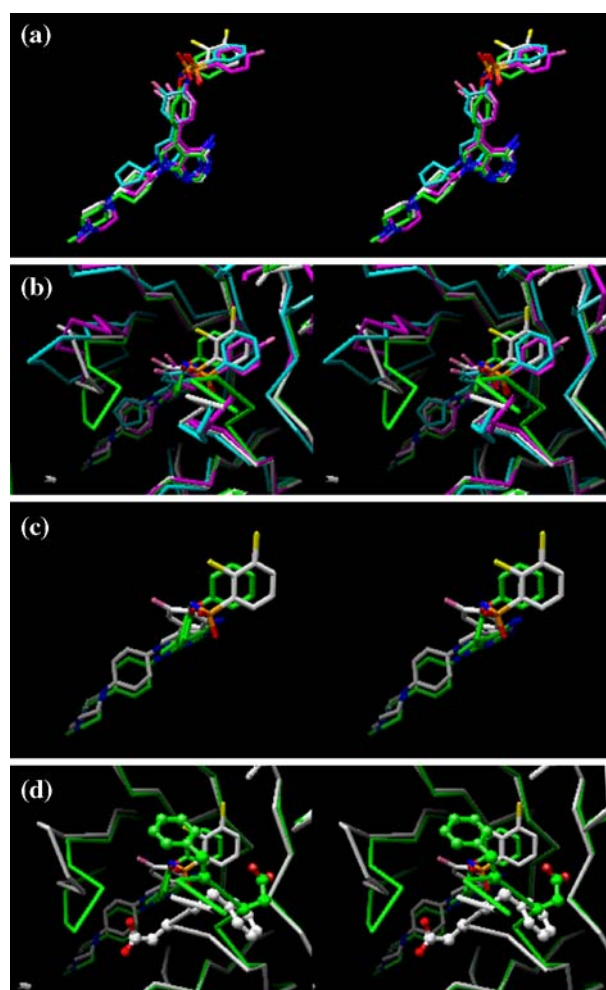


Fig. 2 Stereo-pair views of four pyrrolopyrimidinylamine analogs (structural type I; Compounds 1-1 in cyan, 1-2 in pink, 1-3 in gray, and 1-4 in green) bound to receptor tyrosine kinase Tie-2 from the corresponding superimposed inhibitor-bound Tie-2 crystal structures. (a) Compounds 1-1 through 1-4 have an identical binding mode. (b) Conformation of the glycine-rich loop of Tie-2 shown on the left side is significantly different from each other, even though the superposition of these four inhibitors is very similar. (c) A superposition of Compounds 1-3 and 1-4 bound to Tie-2, whose ligand-bound protein structures are shown in Fig. 2d. (d) The positions of Asp162 and Phe163 residues (of the highly conserved Asp-Phe-Gly motif of Tie-2) from the Compound 1-3-bound (gray) and 1-4-bound Tie-2 (green) are almost opposite in direction. These two residues are shown in a ball-and-stick model

Even though these compounds bind to the same binding site in an identical binding mode, the positions of the binding site residues involved in the ligand–protein interactions are not the same. Therefore, any comparison by SAR and/or QSAR is not under the same condition and may not provide correct information if the ligand-bound protein structures are unknown.

There are three flexible loops in various PK structures discussed in Examples 1–3. They are glycine-rich loop (P-loop), catalytic loop, and activating loop. It is well

documented in the literature that these three loops are flexible and change conformation upon ligand binding [15–17]. They may independently influence the conformation of the proteins upon ligand binding, and provide a flexible binding site for the ligand. Discussions on the subject of the flexible binding sites of other proteins described below, as well as the validity of their structures can be found in the corresponding original references.

Example 2: Pyrimidinone, dihydroquinazolinone, dihydroquinolinone, and phenylpyrazolylphenylmethanone analogs as p38 α MAP kinase inhibitors

Sullivan et al. [18], Stelmach et al. [19] and Fitzgerald et al. [20] reported the X-ray crystal structures of pyrimidinone, dihydroquinazolinone, and dihydroquinolinone

Table 3 Structures of pyrimidinone, dihydroquinazolinone, dihydroquinolinone, and phenylpyrazolylphenylmethanone analogs bound to p38 α MAP kinase and their PDB codes

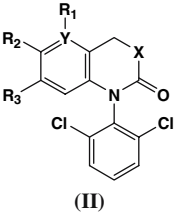
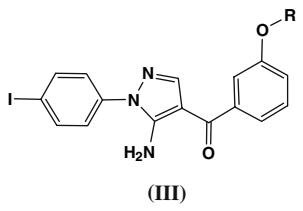
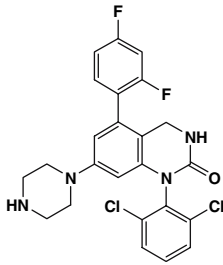
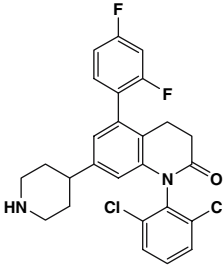
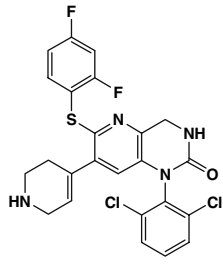
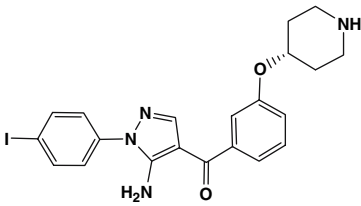
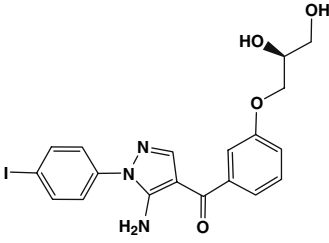
Compd. no.	Structure	PDB/color in Fig. 3
2-1	 <p>(II)</p>  <p>(III)</p>	
2-1		1M7Q (2.40 Å; 0.217) [19]/gray
2-2		1OVE (2.10 Å; 0.195) [20]/green
2-3		1OUY (2.50 Å; 0.226) [20]/magenta

Table 3 continued

Compd. no.	Structure	PDB/color in Fig. 3
3-1		2BAL (2.10 Å; 0.220) [18]/green
3-2		2BAQ (2.80 Å; 0.221) [18]/magenta

The numbers in parentheses after the PDB codes are the resolution and R-values for the corresponding X-ray structural determination

compounds listed in Table 3. The quality of these X-ray structures is reasonable and similar to one another.

Figure 3 shows stereo-pair views of three pyrimidinone, dihydroquinazolinone, and dihydroquinolinone compounds (structural type II, Compounds 2-1 in gray, 2-2 in green, and 2-3 in magenta) and phenylpyrazolyphenylmethanone analogs (structural type III, Compound 3-1 in green and 3-2 in magenta) from the corresponding superimposed inhibitor-bound p38 α MAP kinase crystal structures. Figure 3a shows that binding modes of Compounds 2-1 through 2-3 are essentially identical. Figure 3b shows that the conformation of the glycine-rich loop (Val30–Gly31–Ser32–Gly33–Ala34–Tyr35–Gly36) of Compound 2-2-bound p38 α MAP kinase shown on the left is significantly different from the others. Part of the activating loop (Asp168–Phe169–Gly170–Leu171–Ala172–Arg173) of the same complex shown on the right is also significantly different, even though the superposition of these three inhibitors is very similar.

Figure 3c shows that Compounds 3-1 and 3-2 have an identical binding mode. Figure 3d shows that conformation of the glycine-rich loop (Gly31–Ser32–Gly33–Ala34–Tyr35) of p38 α MAP kinase shown on the left is significantly different from each other. Part of the catalytic- and activating loop (Phe169–Gly170–Leu171–Ala172–Arg173–His174–Thr175–Asp176–Asp177–Glu178–Met179–Thr180–Gly181–Tyr182–Val183–Ala184) shown on the right is also significantly different from each other, even though the superposition of these two inhibitors is very similar. Eleven residues between Ala172 and A184 (14.2 Å) are missing in the protein structure with

Compound 3-1 (colored in yellow) bound and nine residues between Ala172 and Tyr182 (8.5 Å) are missing in the protein structure with Compound 3-2 (colored in cyan) bound. Because of the difference in the glycine-rich loop, as well as the catalytic- and activating loop positions, their amino acid residues of the binding site would naturally be located at different three-dimensional position. For example, Fig. 6d shows that not only is the position of Phe169 of the highly conserved Asp–Phe–Gly motif of the catalytic loop in 2BAL (green, for Compound 3-1) and 2BAK (magenta, for Compound 3-2) different, but their directions as well. In all of the known protein Ser/Thr kinase structures, the Asp–Phe–Gly motif assumes a conformation with the Phe residue buried in a hydrophobic pocket in the groove between the N-terminal and C-terminal lobes of the kinase, which is known to be “DFG-in conformation.” In the structure of the complex with Compound 3-2, the Phe side chain has moved out of the groove to a new position (“DFG-out conformation”) facing towards the solvent.

Even though these compounds bind to the same site in a similar binding mode, the binding site residues involved in the ligand–protein interactions are not the same due to the difference in the flexible binding site residues. Therefore, any comparison by SAR and/or QSAR would not be under an identical condition.

In the above two examples, the binding modes of ligands are the same, whereas the binding sites of the proteins are flexible and thus varied for some ligands. In the next two examples, both the binding modes of ligands and the binding sites of the proteins are different.

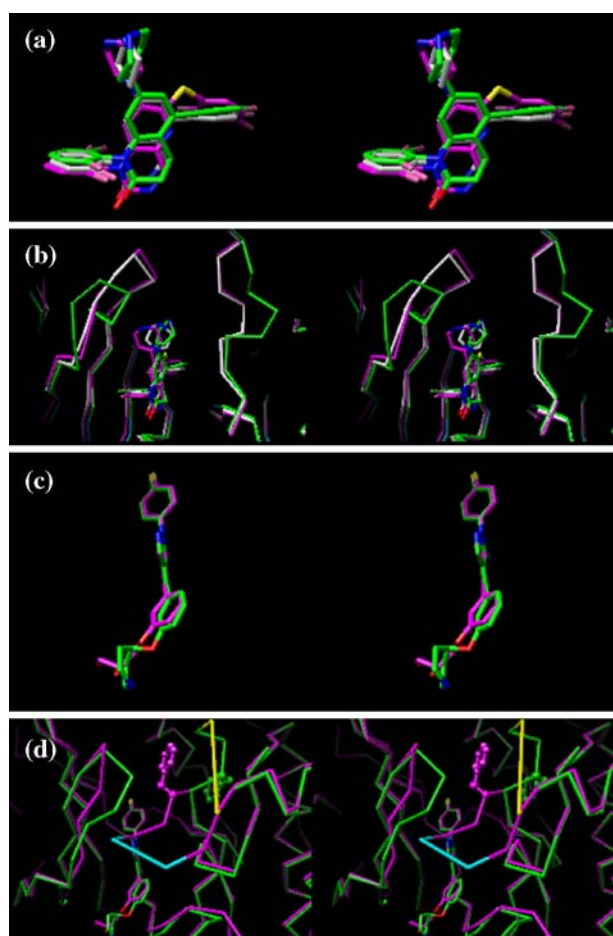


Fig. 3 Stereo-pair views of three pyrimidinone, dihydroquinazolinone, dihydroquinolinone compounds (structural type II, Compounds 2-1 in gray, 2-2 in green, and 2-3 in magenta), and phenylpyrazolylphenylmethanone analogs (structural type III, Compound 3-1 in green and 3-2 in magenta) from the superimposed inhibitor-bound p38 α MAP kinase crystal structures. (a) Compounds 2-1 through 2-3 have an identical binding mode. (b) Conformation of the glycine-rich loop (Val30–Gly31–Ser32–Gly33–Ala34–Tyr35–Gly36) of p38 α MAP kinase shown on the left is significantly different from each other, and part of the activating loop (Asp168–Phe169–Gly170–Leu171–Ala-172–Arg173) shown on the right is also significantly different. (c) Compounds 3-1 and 3-2 have an identical binding mode. (d) Conformation of the glycine-rich loop (Gly31–Ser32–Gly33–Ala34–Tyr35) of p38 α MAP kinase shown on the left is significantly different from each other, and part of the activating loop (Phe169–Gly170–Leu171–Ala-172–Arg173–His174–Thr175–Asp176–Asp177–Glu178–Met179–Thr180–Gly181–Tyr182–Val183–Ala184) shown on the right is also significantly different

Example 3: Sulfonylisoquinoline analogs as cAMP-dependent protein kinase-Rho-kinase surrogate hybrid inhibitors

The activity of cyclic 3',5'-adenosine monophosphate (cAMP)-dependent PKA depends on the presence of cAMP. The enzyme occurs as a 4-membered quaternary structure with two regulatory and two catalytic subunits. In

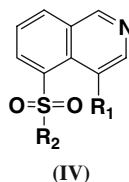
the inactive form of PKA, the regulatory subunits are bound to the active site of the catalytic subunits. When cAMP is present, it binds to the regulatory subunit and causes conformational change releasing and activating the two catalytic subunits. Rho kinase (ROCK) is a serine/threonine kinase and serves as a target protein for small GTP-binding protein Rho [21]. Bonn et al. [22] used site-directed mutagenesis to exchange five amino acids (L49I, V123M, E127D, Q181K, and T183A) in PKA for their counterparts of Rho-kinase and studied binding specificity to the PKA mutants. The quintuple mutants were made to govern the specificity of inhibitors binding to Rho-kinase compared with PKA.

Two compounds (structural type IV) Bonn et al. [22] examined are listed in Table 4. The quality of these X-ray structures is reasonable and similar to one another. Figure 4 displays stereo-pair views of two sulfonylisoquinoline analogs (Compounds 4-1 in orange, 4-2 in gray) from the superimposed inhibitor-bound PKA-Rho-kinase surrogate hybrid crystal structures. Figure 4a shows that Compounds 4-1 and 4-2 have identical binding modes except for the sulfonyldiazepane moiety of the molecules. The sulfonyldiazepane ring of Compound 4-1 is rotated about 90° from the commonly observed binding conformation, as shown with Compound 4-2. The binding mode observed for Compound 4-2 structures (2GNH) is the conformation found for the same compound bound to the threefold mutant (L49I, Q181K, and T183A) model of Rho-kinase [22]. Figure 4b shows how the conformation of the glycine-rich loop of PKA-Rho-kinase surrogate hybrid kinase located above the inhibitor structures is significantly different from each other.

These sulfonylisoquinoline analogs do not bind in similar binding modes, even though their binding sites are identical. Furthermore, the binding site residues involved in the ligand–protein interactions are not the same due to the difference in the flexible binding site residues. Therefore, any comparison by SAR and/or QSAR is not under the same condition.

Example 4: Purine and pyrimidine nucleotide diphosphate compounds as ribonucleotide reductase I substrates

Ribonucleotide reductase (RNR) is an enzyme that catalyzes a crucial step of de novo DNA synthesis by converting nucleoside diphosphates to deoxynucleoside diphosphates. RNRs are divided into three classes based on their cofactor. Class I RNR, found in all eukaryotes, comprises a heterooligomer of α_2 and β_2 subunits. The α subunit called Rnr1 contains the catalytic site, the substrate specificity site, and the activity site. Xu et al. [23, 24] reported the X-ray structures of the eukaryoti α subunit of

Table 4 Structures of sulfonylisoquinoline analogs compounds bound to cAMP-dependent protein kinase (PKA)-Rho-kinase surrogate hybrid and their PDB codes

Compd. no.	Structure	PDB/color in Fig. 4	Compd. no.	Structure	PDB/color in Fig. 4
4-1		2GNI (2.27 Å; 0.192) [22]/orange	4-2		2GNH (2.05 Å; 0.198) [22]/gray

The numbers in parentheses after the PDB codes are the resolution and R-values for the corresponding X-ray structural determination

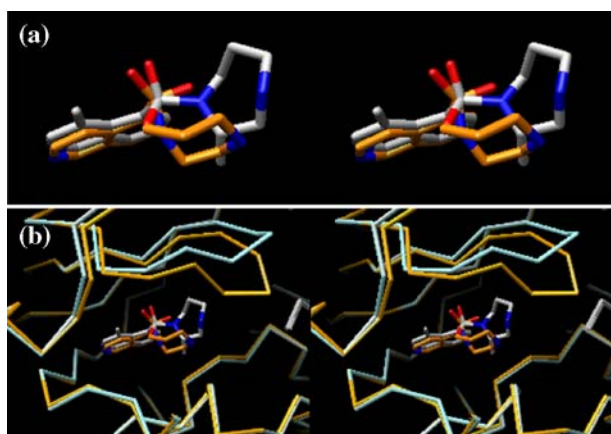


Fig. 4 Stereo-pair views of two sulfonylisoquinoline analogs (structural type IV, Compounds 4-1 in orange, 4-2 in gray) from the superimposed inhibitor-bound cAMP-dependent protein kinase (PKA)-Rho-kinase surrogate hybrid crystal structures. **(a)** Compounds 4-1 and 4-2 have identical binding modes, except the sulfonamidediazepane moiety of the molecules. The position of the sulfonamidediazepane group of Compound 4-1 is different by about 90° from the commonly observed binding conformation, as shown with Compound 4-2. **(b)** Conformation of the glycine-rich loop of PKA-Rho-kinase surrogate hybrid kinase shown above the inhibitor structures is significantly different from each other

RNR from *Saccharomyces cerevisiae* with substrate uridine-5'-diphosphate (UDP; 2CVV), gemcitabine diphosphate (GCQ; 2EUD), cytidine-5'-diphosphate (CDP; 2CVU), adenosine-5'-diphosphate (ADP; 2CVX), and guanosine-5'-diphosphate (GDP; 2CVW). They are listed in Table 5 along with their PDB codes. The quality of Compound 5-1 and 5-3 is not as good as the others. However, it is reasonable and similar to one another.

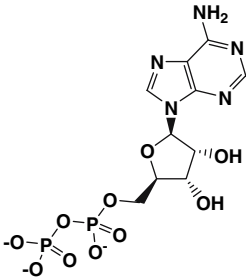
Figure 5 shows stereo-pair views of five nucleotide diphosphate analogs (structural type V, Compounds 5-1 in gray, 5-2 in pink, 5-3 in cyan, 5-4 in yellow, and 5-5 in purple). Figure 5a shows that one cytosine nucleotide analog (Compound 5-1) and the uracil nucleotide analog (Compound 5-3) both have identical binding modes. However, the other cytosine nucleotide analog (Compound 5-2) has a significantly different binding mode. Therefore, this compound may be an outlier in SAR and/or QSAR. Figure 5b shows that the conformation of a part of the protein loop (Asp287–Gln288–Gly289–Gly290–Asn291–Lys292–Arg293–Pro294–Gly295–Ala296) of the RNR structure with Compound 5-2 bound is also significantly different from those RNR structures with Compounds 5-1 or 5-3 bound. Figure 5c shows the superposition of all five nucleotide analogs. The guanine nucleotide analog (Compound 5-4) and the adenine nucleotide analog (Compound 5-5) have identical binding modes as those of Compounds 5-1 and 5-3. Figure 5d shows that even though the binding modes of Compounds 5-4 and 5-5 are similar to those of Compounds 5-1 and 5-3, the protein loop conformation of these are different from one another, as well as from the other three. Since the binding mode of Compound 5-2 is different and the binding site residues involved in the ligand–protein interactions are not the same for all substrates, their SAR and/or QSAR would not be under identical conditions.

The next example is interesting in an additional aspect besides the flexible binding site. Even though the ligands are similar in structure, the protein of interest is from two different sources, and the prevalent conformations of the same protein from these two sources are significantly different.

Table 5 Structures of purine and pyrimidine nucleotide diphosphate compounds bound to ribonucleotide reductase I and their PDB codes

Compd. no.	Structure	PDB/color in Fig. 5
	<p style="text-align: center;">(V)</p>	
5-1		2CVU (2.90 Å; 0.183) [24]/gray
5-2		2EUD (2.30 Å; 0.210) [23]/pink
5-3		2CVV (2.90 Å; 0.184) [24]/cyan
5-4		2CVW (2.40 Å; 0.210) [24]/yellow

Table 5 continued

Compd. no.	Structure	PDB/color in Fig. 5
5-5		2CVX (2.20 Å; 0.215) [24]/purple

The numbers in parentheses after the PDB codes are the resolution and R-values for the corresponding X-ray structural determination

Example 5: Phosphate analogs as triosephosphate isomerase inhibitors

The glycolytic enzyme triosephosphate isomerase (TIM) catalyzes the interconversion of dihydroxyacetone phosphate and glyceraldehydes-3-phosphate. The X-ray structures of TIM revealed “TIM barrel” fold, which is made up of an 8-fold repeat of strand-turn-helix-turn units forming a parallel eight-stranded β -barrel on the inside of the protein surrounded by eight α -helices. Noble et al. [25–27], Verlinde et al. [28], and Parthasarathy et al. [29–31] examined the X-ray crystal structures of phosphate analogs-bound TIM from *Trypanosoma brucei* or *Plasmodium falciparum*, and showed that one of the loop (loop6) involved in different sets of interactions in the “open” and “closed” conformation. Table 6 lists ten phosphate analogs studied along with their PDB codes. The quality of these X-ray structures is reasonable and similar to one another. The qualities of Compound 6-6 and 7-4 (1LZO) are not as good as the other.

Figure 6 shows stereo-pair views of ten phosphate analogs (structural type VI) from the superimposed inhibitor-bound triosephosphate isomerase crystal structures from *Trypanosoma brucei* or *Plasmodium falciparum*. Figure 6a shows that the binding modes of all phosphate analogs (Compounds 6-1 in purple, 6-2 in cyan, 6-3 in dark green, 6-4 in light blue, 6-5 in gray, and 6-6 in yellow) are essentially identical, except for Compound 6-6. Compounds 6-1 through 6-4 are phosphoric acid analogs, and Compound 6-5 and 6-6 are phosphonic acid analogs. Figure 6b shows two loop conformations (1TPD: Val169-Trp170-Ala171-Ile172-Gly173-Thr174-Gly175-Lys176-Val177) shown in ball-and-stick model. Five TIM structures with Compound 6-1, 6-2, 6-3, 6-4 or 6-5 bound are in one conformation (“closed loop” or “swung-in” conformation, colored in orange), and one TIM structure with Compound 6-6 bound (or without any ligand bound) is in the other conformation (“open loop” or “swung-out”

conformation, colored in green). Even though the loop conformation of one phosphonic acid analog (Compound 6-5) is similar to the conformation of the four phosphate analogs, the conformation of the other phosphonic acid analog (Compound 6-6) is different from all the others and similar to the apo (1TPD, [27] not shown) structure. The “swung in” conformation is ideal for catalysis, while the “swung out” conformation is unsuitable for catalysis [26]. Therefore, Compound 6-6 may be an outlier in SAR and/or QSAR within this series.

Figure 6c shows that there are two slightly different binding modes for four phosphate analogs (Compounds 7-1 in yellow, 7-2 in cyan, 7-3 in dark green, and 7-4 in gray or purple), although all of them are very similar: Compounds 7-1, 7-2, and 7-4 (1LZO) in one, and Compounds 7-3 and 7-4 (1LYX) in another. Figure 6d shows that there are two loop conformations as shown in a ball-and-stick model. Four TIM structures with Compound 7-1, 7-2, 7-3, or 7-4 (1LZO) bound (and no ligand bound (1YDV) [32], not shown) are in an open conformation (colored in green), and one TIM structure with Compound 7-4 bound (1LYX) is in a closed conformation (colored in orange). The crystal structures of Compound 7-4 were determined in two different crystal forms: one in orthorhombic ($P2_12_12_1$) and another in monoclinic (C2) form. The catalytic loop adopts the open conformation in orthorhombic crystal (1LZO), and the closed conformation in the C2 form (1LYX).

The majority of the ligand-bound TIM structures adopt the “closed loop” conformation. However, in this latter case, they are mostly in the “open loop” conformation. The observation of both an “open” and a “closed” loop conformations in the ligand-bound state for Compound 7-4 is the first case among the ligand-bound TIM structures. This might show as an outlier in SAR and/or QSAR. Both the “open loop” and partially closed forms of the TIM from *Trypanosoma brucei* was also observed in the sulfate ion-bound structures [33].

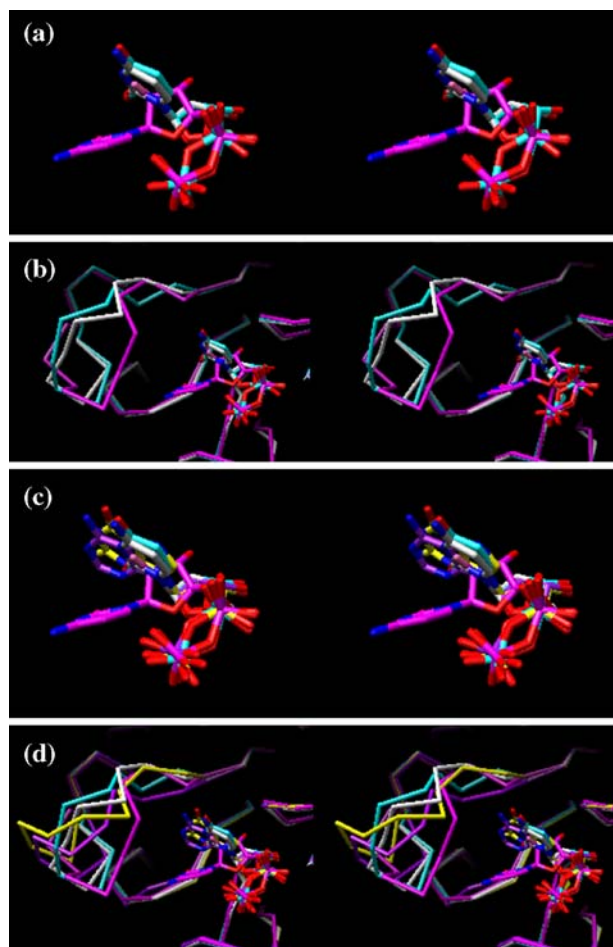


Fig. 5 Stereo-pair views of five nucleotide diphosphate analogs (structural type V, Compounds 5-1 in gray, 5-2 in pink, 5-3 in cyan, 5-4 in yellow, and 5-5 in purple) from the superimposed inhibitor-bound ribonucleotide reductase (RNR) crystal structures. **(a)** One cytosine nucleotide analog (Compound 5-1) and the uracil nucleotide analog (Compound 5-3) essentially have identical binding mode. However, the other cytosine nucleotide analog (Compound 5-2) has a significantly different binding mode. **(b)** The conformation of the protein loop (Asp287–Gln288–Gly289–Gly290–Asn291–Lys292–Arg293–Pro294–Gly295–Ala296) of the RNR structures with Compound 5-2 bound is also significantly different from those of the RNR structures with Compounds 5-1 or 5-3 bound. **(c)** Superposition of all five nucleotide analogs. The guanine nucleotide analog (Compound 5-4) and the adenine nucleotide analog (Compound 5-5) have identical binding mode as those of Compounds 5-1 and 5-3. **(d)** Even though the binding modes of Compounds 5-4 and 5-5 are similar to those of Compounds 5-1 and 5-3, the protein loop conformation of all five are different from one another

The two examples of TIM are quite interesting. Even though both are TIMs, more ligands are found to bind in a closed conformation in the protein structures from *Trypanosoma brucei*. On the other hand, more ligands are found to bind in an open conformation in the corresponding structures from *Plasmodium falciparum*.

Number of ligand-bound X-ray data sets with structurally flexible binding sites

Currently, there are over 40,000 protein structures in the RCSB Protein Data Bank [4]. It is not known how many of them have a flexible binding site due to conformational changes influenced by a bound ligand. In this study, we described the flexible binding sites for six structural data sets associated with five proteins. These are relatively simple and appropriate to demonstrate the flexible binding site for the very structurally close ligands. Flexible binding sites are commonly observed in various protein receptor kinase structures, [17] but also observed in many other proteins [7, 12].

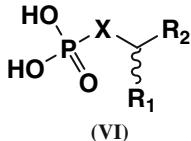
Issues with crystal structures

Several issues need to be considered with regard to the crystal structures as a possible source of outliers in SAR or QSAR. Do the crystal structures represent the real biological binding mode? Is there any uncertainty in the crystal structures? Is it possible for the different binding modes of similar compounds or flexible binding sites to be a result of the differences in the experimental conditions or crystal forms in X-ray crystal determinations? Is the quality similar among different structures? These were all important issues, most of which were included in our previous discussion [3]. The crystal structure may also represent a snapshot of a flexible binding site. Some of the reported X-ray structures may be an artifact due to their crystal packing conditions [32].

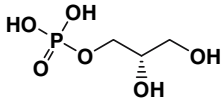
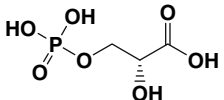
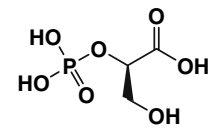
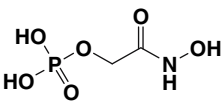
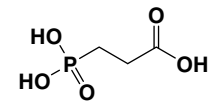
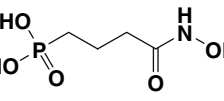
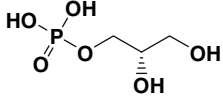
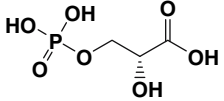
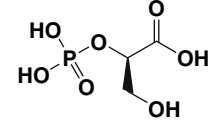
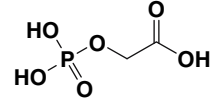
Implications of flexible binding site in SAR and QSAR

The role of protein flexibility is likely to be even greater than the past in drug research. SAR or QSAR studies without ligand-bound protein structural information first assume that analogous compounds bind to the same binding site in the same or similar binding mode. Over four decades, 2D-QSAR studies have been conducted based on this assumption. In many cases this assumption turned out to be correct. When ligand–protein interactions are involved, the observed biological activity is due to the interactions of the ligand with the protein. Therefore, if the orientation of the functional group of one or more ligands is different, or protein changes its binding site conformation upon ligand binding, the descriptors for that functional group are not being calculated to reflect the same ligand–protein interactions. Lack of consideration of binding site flexibility or multiple binding modes can lead to failures not only in understanding important fundamental biological information, but also in developing SAR and/or QSAR.

Table 6 Structures of phosphate analogs bound to triosephosphate isomerase and their PDB codes



(VI)

Compd. no.	Structure	PDB/color in Fig. 6 (ligand–protein)	Compd. no.	Structure	PDB/color in Fig. 6 (ligand–protein)
<i>Trypanosoma brucei</i> triosephosphate isomerase					
6-1		6TIM (2.20 Å; 0.370) [26]/purple–orange	6-2		1IIH (2.20 Å; 0.114) [26]/cyan–orange
6-3		4TIM (2.40 Å; 0.149) [25]/dark green–orange	6-4		1TRD (2.50 Å; 0.147) [27]/light blue–orange
6-5		1IIG (2.60 Å; 0.125) [26]/gray–orange	6-6		1TSI (2.84 Å; 0.115) [28]/yellow–green
<i>Plasmodium falciparum</i> triosephosphate isomerase					
7-1		1M7P (2.40 Å; 0.179) [29]/yellow–green	7-2		1M7O (2.40 Å; 0.183) [29]/cyan–green
7-3		1O5X (1.10 Å; 0.133) [30]/dark green–green	7-4		1LYX (1.90 Å; 0.180) [31]/gray–orange 1LZO (2.80 Å; 0.232) [31]/purple–green

The numbers in parentheses after the PDB codes are the resolution and R-values for the corresponding X-ray structural determination

Possible approaches in 2D QSAR for ligands with different binding modes and flexible binding sites

There are many examples that structurally similar compounds bind to the same binding site in a similar binding mode. However, we observed that even structurally closely related ligands might have different binding modes. In addition, the binding site may not stay rigid, but change its conformation upon ligand binding. Then, how can one do SAR and/or QSAR in a lead optimization process?

If one wants to be sure that all the ligands included have the same binding mode or the binding site does not change, one has to determine the binding mode of each ligand by X-ray or neutron diffraction crystallography or even by NMR techniques. However, it is still not practical in most places to determine structures for every ligand-bound protein structure. Therefore, when multiple

binding modes or flexible binding sites are suspected for structurally close analogs, one may examine both 2D and 3D-QSAR (such as Comparative Molecular Field Analysis, CoMFA) [34–36] and compare the results to see whether they indicate similar SAR. While 2D-QSAR studies assume the same binding mode, 3D-QSAR studies include ligands with different binding modes. However, when a flexible binding site is suspected, even a 3D-QSAR may not be helpful, because the receptor site environment for each molecule is different. In such a case, one may have to do individual analysis including conformational analysis, or try to identify potential outliers. Of course, the traditional way of doing QSAR will certainly continue if one desires, but one has to pay close attention to the outliers more closely considering the possibilities of multiple binding modes and/or flexible binding sites.

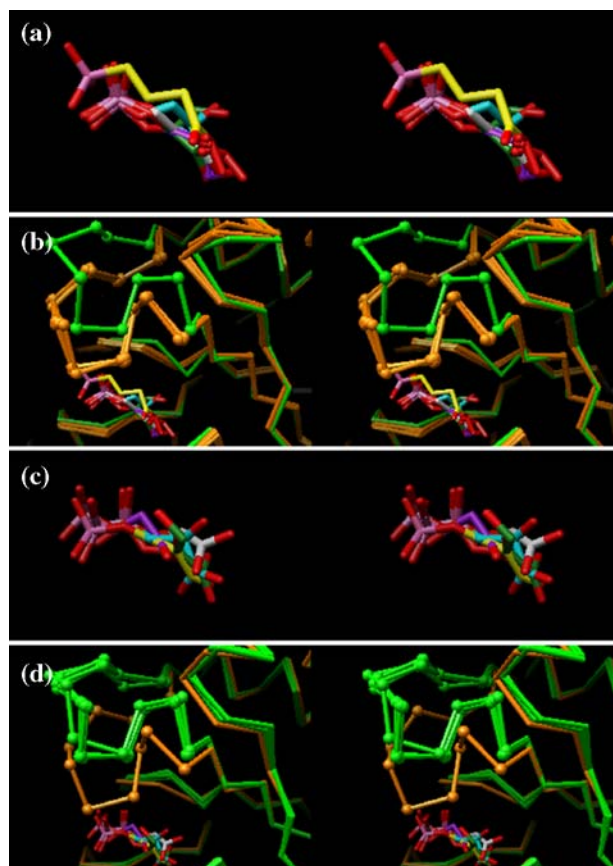


Fig. 6 Stereo-pair views of ten phosphoric and phosphonic acid analogs from the superimposed inhibitor-bound triosephosphate isomerase crystal structures from *Trypanosoma brucei* or *Plasmodium falciparum*. (a) The binding modes of all analogs (structural type VI, Compounds 6-1 in purple, 6-2 in cyan, 6-3 in dark green, 6-4 in light blue, 6-5 in gray, and 6-6 in yellow) are essentially identical except Compound 6-6. (b) There are two loop conformations (1TPD: Val169-Trp170-Ala171-Ile172-Gly173-Thr174-Gly175-Lys176-Val177) shown in a ball-and-stick model. Five TIM structures with Compound 6-1, 6-2, 6-3, 6-4, or 6-5 bound are one conformation and colored in orange, and one TIM structure with Compound 6-6 bound (or without any ligand bound, 1TPD [27], not shown) is the other conformation and colored in green. (Noble et al. [26] reported that Compound 6-2 has a “swung out” conformation.) (c) Four phosphate analogs (structural type VI, Compounds 7-1 in yellow, 7-2 in cyan, 7-3 in dark green, and 7-4 in gray or purple) are in two slightly different binding modes: Compounds 7-1, 7-2, and 7-4 (1LZO) in one, and Compounds 7-3 and 7-4 (1LYX) in another. (d) There are two loop conformations as shown in a ball-and-stick model. Four TIM structures with Compound 7-1, 7-2, 7-3, or 7-4 (1LZO) bound (and apo (1YDV) [32], not shown) are in one conformation (colored in green), and one TIM structure with Compound 7-4 bound (1LYX) is in the other conformation (colored in orange)

Conclusions

Several types of flexible binding sites have been observed. Some of them resulted from loop motion, while others from domain motion. Between the essentially identical binding site structures with multiple ligand binding modes and the

different binding site structures due to the flexibility of the binding site with essentially an identical binding mode of ligands, there are numerous cases in between. Some examples that we have studied so far are cases with both multiple binding modes and different binding sites due to flexibility of the binding site. Of course, only the binding site residues may move in some cases.

Such flexible binding sites for some of the analogs presented here do not actually provide indisputable evidence for the outliers observed in QSARs in Table 1. However, there is no doubt that compounds bound to one conformation of flexible binding site would not give the same SAR and QSAR or fit to the QSAR generated from the compounds bound to a different conformation of flexible binding site. Different binding sites of flexible binding sites observed in X-ray crystal structures for analogous compounds provide the chance for it to be a source of outliers observed in SAR or QSAR studies.

Flexible binding sites and multiple binding modes presented here remind us that the first assumption in SAR or QSAR that analogous compounds bind to the same binding site in the same or similar binding mode should really be the first assumption. One should soon readjust the assumption based on the observed SAR or QSAR in the continuing process of lead optimization. Outliers can be very important and may provide new opportunities in drug discovery research. Thus, they should not be treated lightly.

Acknowledgments The author expresses sincere gratitude to both Professor Corwin Hansch and Dr. Albert Leo for their generous permission to use their C-QSAR database and the BioByte program Bio-Loom. The author would also like to thank Mrs. Angela Mun for proofreading this manuscript.

References

- Kurup A (2003) J Comput Aided Mol Des 17:187–196
- BioByte 201 W. 4th St., #204, Claremont, CA 91711-4707. clogp@biobyte.com. 909-624-5992
- Kim KH (2007) J Comput Aided Mol Des 21:63–86
- Berman HM, Westbrook J, Feng Z, Weissig H, Shindyalov IN, Bourne PE (2000) Nucleic Acids Res 28:235–242
- Thompson J, Jeanmougin F (2003) Clustal W multiple sequence alignment program (Version 1.83, June 2003)
- Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE (2004) J Comput Chem 25:1605–1612
- Teague SJ (2003) Nat Rev Drug Discov 2:527–541
- Davis AM (1999) Angew Chem Int Edit 38:736–749
- Gutteridge A, Thornton J (2005) J Mol Biol 346:21–28
- Najmanovich R, Kuttner J, Sobolev V, Edelman M (2000) Proteins: Struct Func Genet 39:261–268
- Li W, Liang S, Wang R, Lai L, Han Y (1999) Protein Eng 12:1075–1086
- Gunasekaran K, Nussinov R (2007) J Mol Biol 365:257–273
- Gerstein M, Lesk AM, Chothia C (1994) Biochemistry 33:6739–6749

14. Bump NJ, Arnold LD, Dixon RW, Hoeffken HW, Allen K, Bellamacina C (2001) Method of identifying inhibitors of Tie-2. WO 01/72778, October 4, 2001
15. Hubbard SR (2000) *Ann Rev Biochem* 69:373–398
16. Noble MEM (2004) *Science* 303:1800–1805
17. Huse M (2002) *Cell* 109:275–282
18. Sullivan JE, Holdgate GA, Campbell D, Timms D, Gerhardt S, Breed J, Breeze AL, Bermingham A, Pauptit RA, Norman RA, Embrey KJ, Read J, Vanscyoc WS, Ward WH (2005) *Biochemistry* 44:16475–16490
19. Stelmach JE, Liu L, Patel SB, Pivnichny JV, Scapin G, Singh S, Hop CECA, Wang Z, Cameron PM, Nichols EA, O’Keefe SJ, O’Neill EA, Schmatz DM, Schwartz CD, Thompson CM, Zaller DM, Doherty JB (2003) *Bioorg Med Chem Lett* 13:277–280
20. Fitzgerald CE, Patel SB, Becker JW, Cameron PM, Zaller DM, Pikounis VB, O’Keefe SJ, Scapin G (2003) *Nat Struct Biol* 10:764–769
21. Yamaguchi H, Kasa M, Amano M, Kaibuchi K, Hakoshima T (2006) *Structure* 14:589–600
22. Bonn S, Herrero S, Breitenlechner CB, Erlbruch A, Lehmann W, Engh RA, Gassel M, Bossemeyer D (2006) *J Biol Chem* 281:24818–24830
23. Xu H, Faber C, Uchiki T, Racca J, Dealwis C (2006) *Proc Natl Acad Sci USA* 103:4028–4033
24. Xu H, Faber C, Uchiki T, Fairman JW, Racca J, Dealwis C (2006) *Proc Natl Acad Sci* 103:4022–4027
25. Noble ME, Verlinde CL, Groendijk H, Kalk KH, Wierenga RK, Hol WG (1991) *J Med Chem* 34:2709–2718
26. Noble ME, Wierenga RK, Lambeir AM, Opperdoes FR, Thunnissen AM, Kalk KH, Groendijk H, Hol WG (1991) *Proteins* 10:50–69
27. Noble ME, Zeelen JP, Wierenga RK (1993) *Proteins* 16:311–326
28. Verlinde CL, Witmans CJ, Pijning T, Kalk KH, Hol WG, Callens M, Opperdoes FR (1992) *Protein Sci* 1:1578–1584
29. Parthasarathy S, Balaram H, Balaram P, Murthy MRN (2002) *Acta Crystallogr Sect D* 58:1992–2000
30. Parthasarathy S, Eaazhisai K, Balaram H, Balaram P, Murthy MR (2003) *J Biol Chem* 278:52461–52470
31. Parthasarathy S, Ravindra G, Balaram H, Balaram P, Murthy MR (2002) *Biochemistry* 41:13178–13188
32. Velanker SS, Ray SS, Gokhale RS, Suma S, Balaram H, Balaram P, Murthy MR (1997) *Structure* 5:751–761
33. Wierenga RK, Noble MEM, Vriend G, Nauche S, Hol WGJ (1991) *J Mol Biol* 220:995–1015
34. Kim KH (1995) Comparative molecular field analysis. In: Dean PM (ed) *Molecular similarity in drug design*. London, Chapman & Hall, pp 291–331
35. Kim KH, Greco G, Novellino E (1998) *Perspect Drug Discov Des* 12–14:233–255
36. Martin YC, Kim KH, Lin CT (1996) *Comparative molecular field analysis: CoMFA vol 1*. JAI Press, Greenwich, CT pp 1–52