Chemistry & Biology
Article

Molecular Similarity Analysis Uncovers Heterogeneous Structure-Activity Relationships and Variable Activity Landscapes

Lisa Peltason[1] and Jürgen Bajorath[1,*]
[1] Department of Life Science Informatics, B-IT, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany
*Correspondence: bajorath@bit.uni-bonn.de
DOI 10.1016/j.chembiol.2007.03.011

SUMMARY

We systematically compare X-ray structures of inhibitor complexes of four well-known enzymes and correlate two- and three-dimensional (2D and 3D) similarity of inhibitors with their potency. The analysis reveals the presence of unexpected systematic relationships between molecular similarity and potency. These findings explain why apparently inconsistent structure-activity relationships (SARs) can coexist in different targets, and they have general implications for compound screening and optimization. The results suggest that (1) even for active sites with significant binding constraints, there is a high probability that structurally diverse ligands with similar activity can be identified, (2) different types of SARs are not mutually exclusive, and (3) the chemical nature of ligands is of comparable importance for SARs as the features of active sites. These insights aid in the understanding of target-specific SARs and their intrinsic degree of variability.

INTRODUCTION

Understanding the relationship between the similarity and biological specificity of small molecules is of paramount importance for drug design and chemical biology [1–5]. Specific binding of small molecules to target proteins is generally characterized by a high degree of molecular complementarity including specific interactions and shape [4]. How these binding characteristics might generally affect the nature of ligand structure-activity relationships (SARs) is unknown. Open questions include, for example: Do structural features of target sites ''dictate'' SARs? Are different types of SARs mutually exclusive? Is ligand similarity related to SAR characteristics?

Importantly, the study of SARs is complicated by a widely recognized conundrum: on the one hand, small structural modifications of active molecules often have dramatic effects on selectivity and/or potency [6]. Therefore, we call the underlying SARs ''discontinuous.'' On the other hand, a spectrum of similar molecules is often found to have similar activity [5], and, in addition, increasingly structurally diverse compounds can be active against the same target [5, 7]. Accordingly, the underlying SARs are termed ''continuous.'' Discontinuous SARs are explored in chemical lead optimization [6], whereas continuous SARs are investigated in molecular-similarity analysis [5], small-molecule-based virtual screening [7], or ''scaffold hopping'' [8]. Although principal differences between SAR characteristics are long known [5], potential relationships between continuous and discontinuous SARs have remained largely unclear.

In our analysis, we have evaluated SAR characteristics in detail with the aid of experimental structural and binding data and different molecular representations for the evaluation of similarity. We have compared the crystal structures of complexes of well-known enzymes with different inhibitors and correlated the 2D and 3D similarity of bound inhibitors with their potency. The analysis reveals systematic and, in part, unexpected relationships between similarity and potency and the coexistence of different target-specific SARs within an activity landscape. This helps to rationalize why molecular-similarity methods often succeed in identifying novel active molecules [5, 7]. Furthermore, variations in 2D structure do not always constrain the ability of ligands to adopt diverse binding modes, which has important implications for drug design.

RESULTS AND DISCUSSION

On the basis of a survey of the PDBbind database [9, 10], we selected four enzymes for our analysis: two serine proteases, elastase and coagulation factor Xa; carbonic anhydrase II; and ribonuclease A. The wealth of structural data available for these enzymes is a consequence of the fact that they are long-established targets. In addition, these test cases represent active sites of distinctly different architectures and chemical features. Figure 1 shows the enzyme structures and active-site regions, and Table S1 (see the Supplemental Data available with this article online) summarizes the structural data used for our analysis. We have systematically compared the 2D similarity of

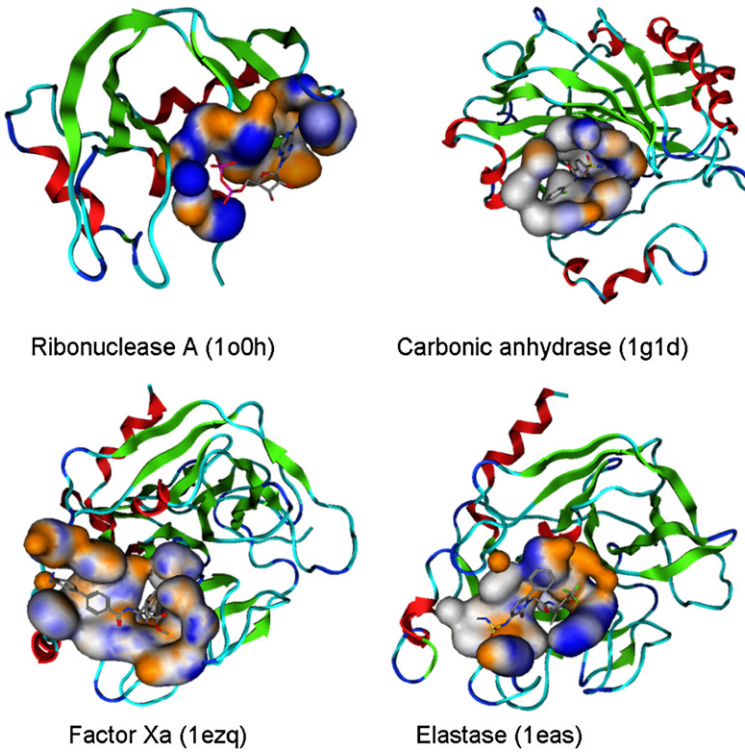

Figure 1. Structures of Enzyme Targets and Their Active Site Regions

For each of the four enzymes studied here, a representative X-ray structure of an inhibitor complex is shown (with its PDB ID code). Active sites are rendered in solid surface representations and are colored by partial charge distributions (gold, negative; blue, positive). When atom coloring is used for inhibitors, the same color scheme is applied in all representations: carbon, gray; oxygen, red; nitrogen, blue; sulfur, yellow; phosphorus, magenta; halogens, green.

inhibitors with their 3D similarity (determined by binding conformation, position, and orientation), quantitatively assessed similarity relationships, and related these relationships to differences in compound potency. There are no absolute measures of molecular similarity, and any evaluation of similarity depends on the molecular representations that are employed (as described in Experimental Procedures). However, similarity relationships discussed in our study can also be qualitatively appreciated by comparing the 2D molecular graphs and 3D structural views that we present.

The active sites of ribonuclease A and carbonic anhydrase II pose severe structural constraints on ligand binding. Ribonuclease A cleaves single-stranded RNA and accommodates a phosphate group in a positively charged phosphate-binding pocket (Figure S1). Filling this pocket and compensating the positively charged residues is a prerequisite for inhibitor binding, which makes this enzyme a model system for SARs that are discontinuous in nature and largely determined by the presence of an "activity cliff" [11]. Accordingly, known inhibitors are nucleotide analogs with one or more phosphate group and typically have very similar structures (Figure 2A) that have Tanimoto similarity between MACCS structural key representations greater than 0.75 (for an explanation, see Experimental Procedures). Although these inhibitors have very similar 2D structures, there are significant 3D variations (Figure 2A) because inhibitors containing different bases bind in different conformations and orientations. In fact, as illustrated in Figure 2B, similar structures can bind very differently as long as the phosphate group constraint is satisfied. Furthermore, there is little correlation between structural similarity and compound potency. Similar structures have different potency levels irrespective of whether their binding modes are similar or not. Thus, although the underlying SAR looks very simple at the 2D level, there is significant variation among ribonuclease A inhibitor-binding modes. Moreover, inhibitors with nearly identical binding conformations can differ by up to three orders of magnitude in potency, as illustrated in Figure 2C. As a prototype for discontinuous SARs, ribonuclease A is surprisingly flexible in its accommodation of different binding modes.

In carbonic anhydrase II, the need to coordinate a catalytically important zinc cation within the active site presents the major constraint for inhibitor binding (Figure S2), and sulfonamide groups are a hallmark of carbonic anhydrase inhibitors, contributing several orders of magnitude to potency (Figure S2). On the basis of these characteristics, one might expect that carbonic anhydrase inhibitors share discontinuous SARs, similar to ribonuclease A. However, as shown in Figure 3, continuous SARs exist among sulfonamide derivatives. There is significant correlation between the 2D and 3D similarity (Figures 3A and 3B). Furthermore, dissimilar structures have the greatest differences in 3D similarity and potency, whereas similar compounds bind in a similar manner and with comparable potency (Figures 3A and 3C). Thus, in this case, continuous SARs exist proximal to an "activity cliff," a boundary provided by the sulfonamide constraint.

Factor Xa and elastase present examples of active sites with less stringent requirements for inhibitor binding than ribonuclease A or carbonic anhydrase II. As shown in Figure 4A, the majority of factor Xa inhibitors are related

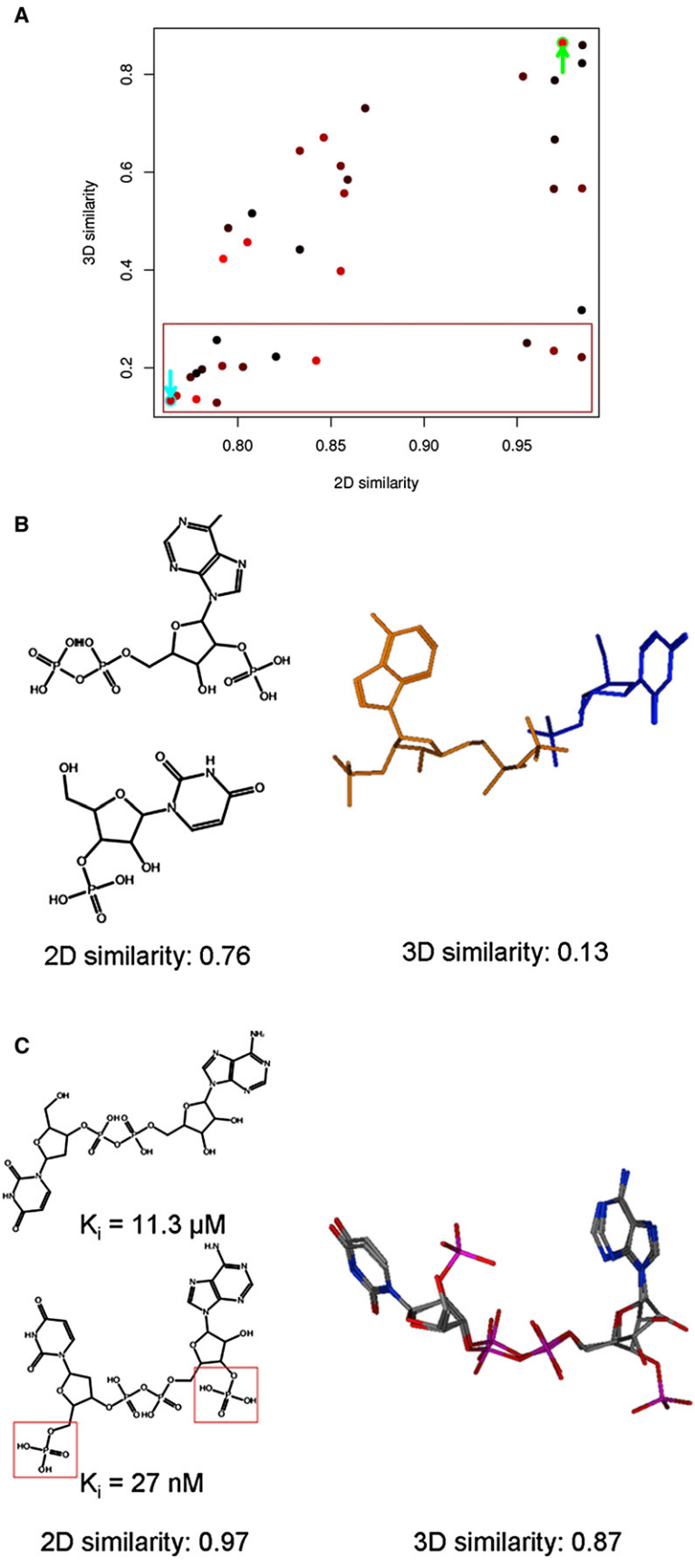

Figure 2. Ribonuclease A Inhibitors

(A) Comparison of 2D and 3D similarity. Each dot represents values of a pairwise comparison of two inhibitors. Data points are color-coded according to potency differences by using a continuous spectrum from black (smallest potency difference) to red (largest potency difference). The blue and green arrows identify the inhibitor pairs shown in (B) and (C), respectively. The red box indicates inhibitors with similar structures that adopt very different binding modes. The correlation coefficient of 2D and 3D similarity is 0.58.

(B) Examples of similar inhibitors that bind very differently in the ribonuclease A active site. On the left side, the 2D structure of the inhibitors is shown. On the right, the same inhibitors are shown in their relative binding conformations and orientations after optimal superposition of the enzyme α carbon atoms.

(C) Examples of compounds with very similar binding modes but dramatic differences in potency.

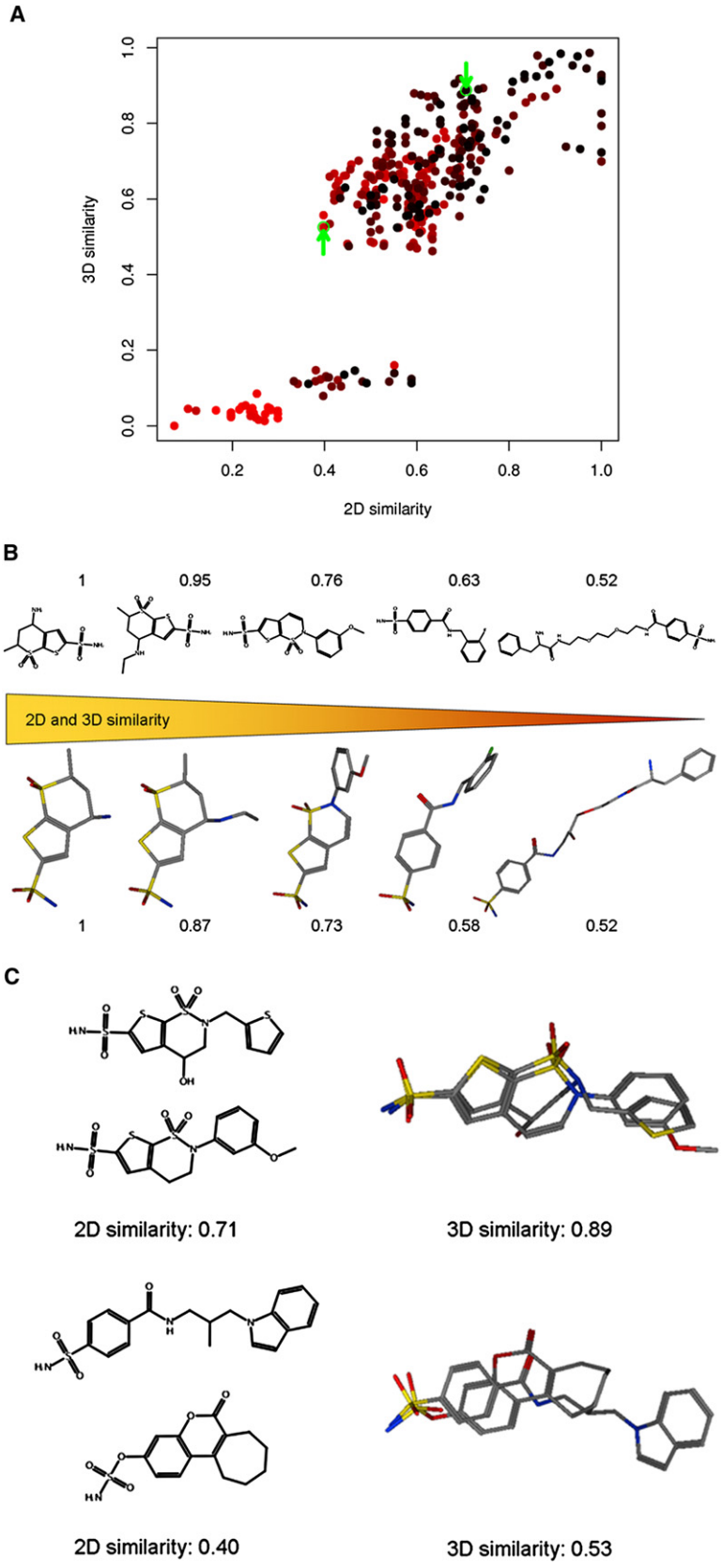

Figure 3. Carbonic Anhydrase Inhibitors

(A) Comparison of 2D and 3D similarity (represented according to Figure 2A). Compounds with the lowest 2D and 3D similarity have the greatest differences in potency (lower left). The green arrows identify the inhibitor pairs depicted in (C). The correlation coefficient between 2D and 3D similarity is 0.79.

(B) Direct correlation between 2D and 3D similarity among sulfonamides. 2D and 3D similarity values are reported for pairwise comparisons by using the inhibitor on the left as the reference compound. The 2D and 3D similarity to the reference compound decreases from left to right, as indicated by the yellow wedge.

(C) Examples of sulfonamide inhibitors with significant (top) or limited (bottom) 2D/3D similarity.

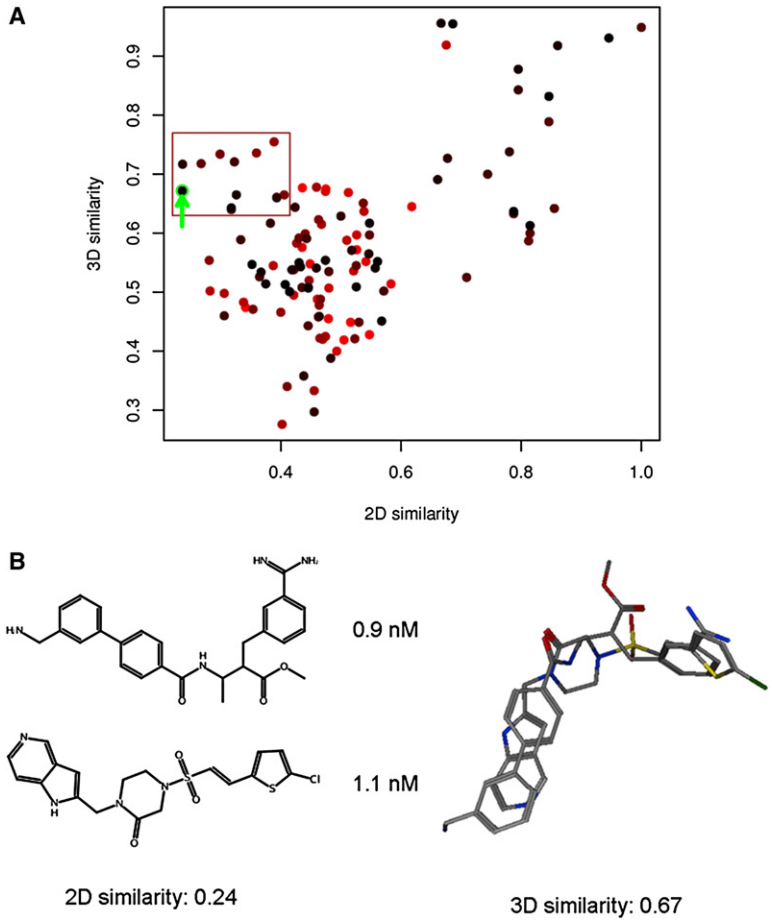

Figure 4. Factor Xa Inhibitors

(A) Comparison of 2D and 3D similarity. The red box indicates inhibitors that adopt similar binding modes despite limited structural similarity. The arrow indicates the inhibitor pair shown in (B). The correlation coefficient between 2D and 3D similarity is 0.47.

(B) Examples of inhibitors with low 2D but distinct 3D similarity.

by a continuous SAR. In contrast to ribonuclease A, factor Xa inhibitors display significant structural diversity. There is detectable correlation between 2D and 3D similarity, and most similar 2D structures bind very similarly and with comparable potency. By contrast, the largest differences in potency are observed among inhibitors with limited 2D and 3D similarity. A characteristic feature of factor Xa inhibitors is that structures with very low 2D similarity (representing distinct chemical scaffolds) can also adopt very similar binding modes, as illustrated in Figure 4B. For these inhibitors, matching the shape of the active site and forming only a few key interactions are of particular importance for binding (Figure S3).

Next, we studied elastase. At first glance, there is little correlation between the 2D and 3D similarity of inhibitors (Figure 5A). However, we identified subsets of elastase inhibitors for which 2D and 3D similarity is either strongly or inversely correlated (Figure 5B). Selected elastase-inhibitor complexes are shown in Figure 6. For a subset of inhibitors consisting of peptide derivatives and other compounds, strong correlation between structural and binding similarity is observed, and compound potency is found to increase with decreasing 2D/3D similarity (Figure 5C). Thus, if we consider the most potent compound as a reference point, gradual structural departures from a preferred inhibitor are accompanied by a gradual loss in potency, which represents a prime example of a continuous SAR. On the other hand, for a series of trifluoro-acetyl (TFA)-dipeptide-anilides with overall comparable potency, 2D and 3D similarities inversely correlate (Figure 5D). This means that within this series of inhibitors, decreasingly similar compounds adopt increasingly similar binding modes, which represents a different type of a continuous SAR. How can these observations be rationalized? As shown in Figure 6A, a series of TFA-dipepetide-anilides adopt three distinct binding modes within the active site of elastase. However, each of these binding modes can be adopted by 2D diverse inhibitors that present their functional groups in spatially corresponding positions (Figure 6B). Thus, in the case of elastase, different continuous SARs can be distinguished. Structures of elastase-inhibitor complexes that represent a series of inhibitors with inversely or directly correlated 2D/3D similarity are shown in Figure S4.

In summary, inhibitor binding to ribonuclease A is governed by discontinuous SARs, albeit with a remarkable degree of 3D variability. By contrast, factor Xa inhibitors present a prototypic example of continuous SARs. In carbonic anhydrase II, significant correlation between 2D and 3D similarity within the boundaries of a structural constraint is observed, revealing the presence of a continuous SAR within the limits of a discontinuous one. This situation

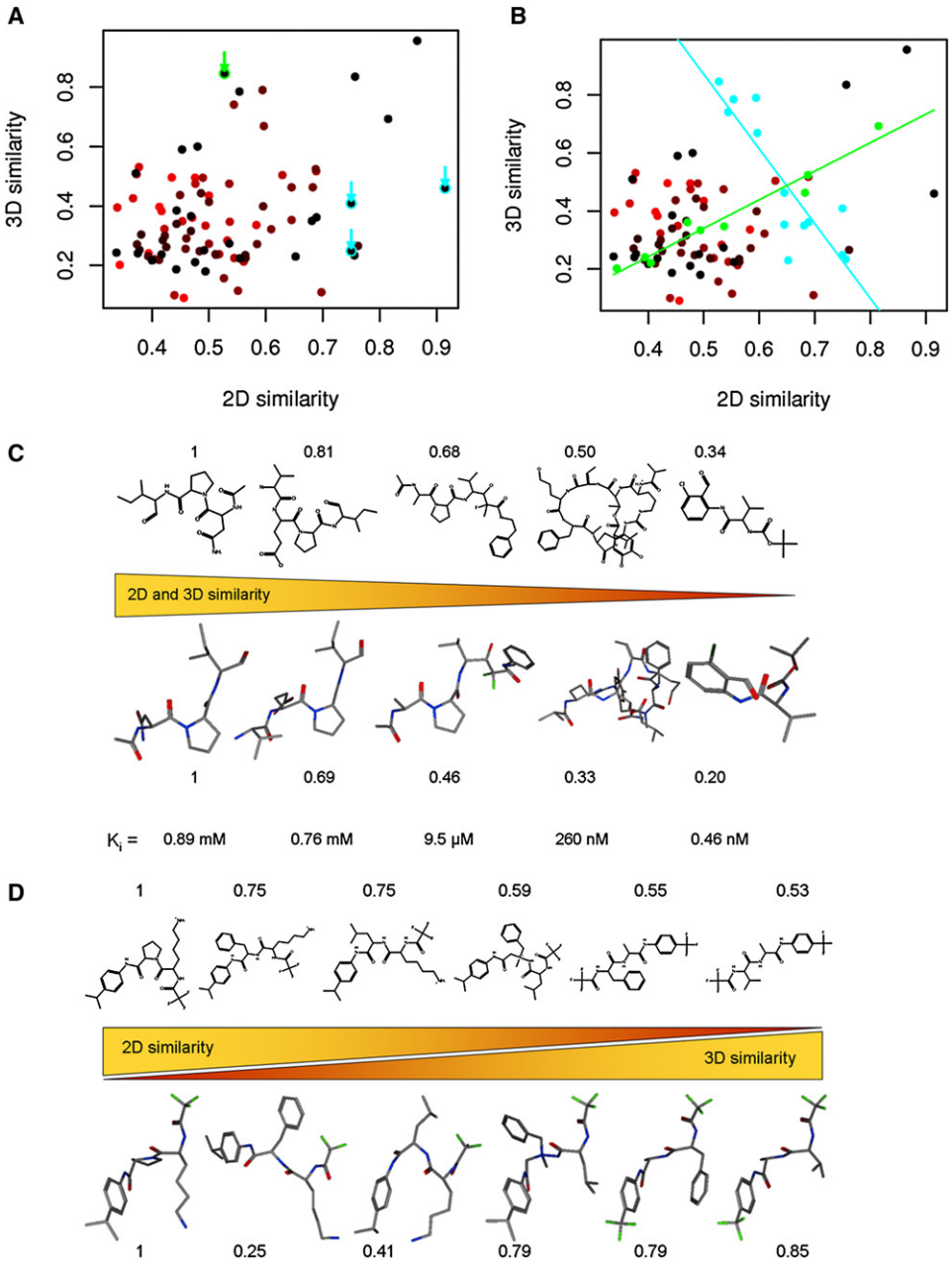

Figure 5. Elastase Inhibitors

(A) Comparison of 2D and 3D similarity. The overall correlation between 2D and 3D similarity is 0.31. The blue and green arrows identify the inhibitors shown in Figures 6A and 6B, respectively.

(B) Identification of subsets of elastase inhibitors in (A) with strong direct (green) or inverse (light blue) correlation between 2D and 3D similarity. The green dots refer to the subset shown in (C), and the blue dots refer to the subset in (D). The molecules in these subsets were selected because they are representative data points in linear models of direct and inverse 2D/3D similarity correlation. Correlation coefficients for direct and inverse correlation are 0.97 and −0.87, respectively.

(C) Direct correlation between 2D and 3D similarity in a subset of elastase inhibitors (represented as in Figure 3B). The most potent compounds have the lowest 2D and 3D similarity to others.

(D) Inverse correlation between 2D and 3D similarity in another subset of elastase inhibitors (with comparable potency).

is representative of heterogeneous SARs and likely applies to many different binding sites. Finally, in the case of elastase, different continuous SARs characterize different series of elastase inhibitors. For one series, there is significant correlation between 2D and 3D similarity. By contrast, for another series of inhibitors, 2D and 3D

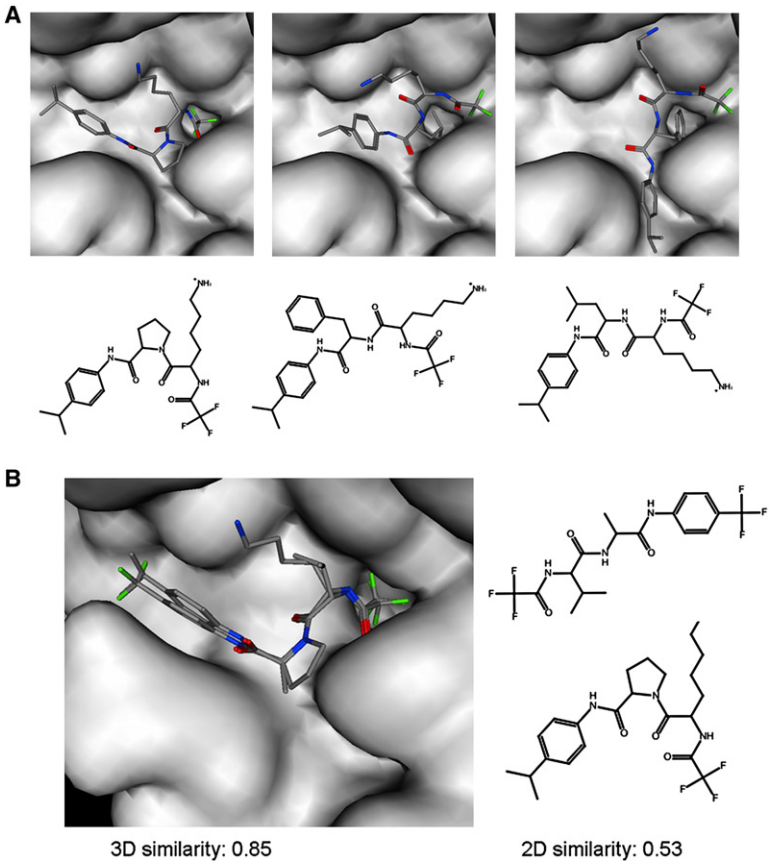

Figure 6. Alternative Binding Modes in Elastase

(A) Shown are three structurally similar inhibitors that adopt distinct binding modes in the active site of elastase.

(B) Each of these binding modes is also shared by inhibitors with limited 2D similarity. As an example, two inhibitors are shown that adopt the binding mode shown on the left in (A).

similarity is inversely correlated. However, within this series, 2D similar inhibitors can adopt three distinct binding modes, each of which is also shared by diverse ligands. These findings revise current views that similar ligands generally bind in a very similar way to the same target [12] and make it possible to directly relate aspects of molecular similarity to SAR characteristics.

Taken together, our results reveal systematic trends in the comparison of 2D and 3D similarity and show that even severe constraints on binding to active sites permit significant variability of compound-binding modes and the coexistence of discontinuous and continuous SARs. Thus, we demonstrate the heterogeneous nature of SARs in a target-specific activity landscape. Such SAR characteristics were previously proposed to be of crucial importance for the successful application of molecular-similarity methods [5]. Moreover, continuous SARs are observed for distinct active sites, and different continuous SARs that are dependent on the features of ligands can coexist in an enzyme. The picture that emerges from our analysis of four "classic" enzyme targets is that different SARs are not mutually exclusive and are more heterogeneous in nature than often thought. Even in rugged activity landscapes, continuous regions exist. These findings imply that different chemical scaffolds with similar activity can likely be identified for many different protein targets by searching for continuous regions of activity landscapes through experimental or computational compound screening.

SIGNIFICANCE

The study of small-molecule SARs is one of the major topics in medicinal chemistry, drug design, and chemical biology. However, the relationship between continuous and discontinuous SARs is currently not well understood. Here, we have analyzed X-ray structures of enzyme-inhibitor complexes and corresponding binding data in order to systematically correlate the 2D and 3D similarity and potency of ligands. Given the molecular representations we have chosen, we find that 2D and 3D similarity and compound potency can be related in previously unobserved ways. These results provide evidence for the presence of multiple and heterogeneous SARs within active sites and activity landscapes. These findings suggest that it should be possible to identify small molecules with diverse structures but similar activity for many enzymes and probably other target proteins.

EXPERIMENTAL PROCEDURES

Binding conformations and orientations of inhibitors were compared after optimal superposition of α carbon atoms of all enzyme structures (Table S1) by using the sequence/structure alignment function of the Molecular Operating Environment (MOE, Chemical Computing Group, Inc.; http://www.chemcomp.com/). For the analysis of 2D and 3D molecular-similarity relationships, molecular descriptors and representations are critical parameters. Our calculation of 2D similarity

focuses on conventional structural fragment-type descriptors. As a measure of 2D similarity, the Tanimoto coefficient (Tc) [13] was calculated (see below) using a fingerprint consisting of the publicly available set of 166 MACCS structural keys (MDL Elsevier; http://www.mdl.com/). As a measure of 3D similarity, the normalized overlap of atomic property density functions [14] was calculated (details are provided below), which takes both conformational and positional differences into account. Both 2D and 3D similarity values range from 0 to 1. X-ray structures were taken from PDBbind [9, 10] or from the RCSB Protein Data Bank [15]. For inhibitors, we only considered experimentally determined conformations and binding modes, not modeled structures. Potency values (K_i or K_D) for every crystallized inhibitor were taken from PDBbind or original references and were transformed into logarithms to the base of 10. Structural representations were generated with MOE.

Calculation of 2D Similarity

The 2D similarity of two compounds was computed by means of the Tanimoto coefficient by using the MACCS fingerprints as implemented in MOE. This publicly available fingerprint contains 166 bits indicating the presence of specified structural fragments in the molecular graph representation. The Tc is a similarity measure for fingerprint overlap and counts the number of bits common to two binary fingerprints with respect to the total number of bits that are set on in each fingerprint. The Tc for two binary fingerprint representations, *A* and *B*, is calculated as follows:

$$Tc(A,B) = \frac{N_{AB}}{N_A + N_B - N_{AB}}, \quad (1)$$

where N_{AB} is the number of bits that are set on in both fingerprints, and N_A and N_B refer to the number of bits that are set on in *A* and *B*, respectively. Given this formulation, identical fingerprints have a Tc value of 1, whereas nonoverlapping fingerprints are assigned a Tc value of 0.

Calculation of 3D Similarity

For the comparison of the conformation and spatial position of two bound ligands, a property density function was defined for each molecule, and the normalized overlap of both functions was calculated as a measure of 3D molecular similarity.

First, a common reference frame was established by superposition of the protein α carbon atoms by using the sequence/structure alignment function of MOE. Then, a property density function for the coordinates of each ligand was defined and calculated as follows. Each atom is represented by a spherically symmetric Gaussian density function centered at the position of the atom nucleus; width is determined by the van der Waals atom radius. The density function of a molecule is then defined with respect to specified atomic properties by the weighted sum of the density functions of all atoms present in the molecule. The atomic Gaussians are weighted with respect to selected atom properties. We weight the atomic property density by 1 if the atom has a specified property, and by 0 if not. The four selected properties are aromaticity, hydrogen-bond acceptor potential, hydrogen-bond donor potential, and hydrophobicity.

The overlap of the property density functions of two molecules is the sum of the respective property density functions, which is again a Gaussian (Figure S5 provides a graphical illustration of the approach):

$$F(X,Y) = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{w_i^p w_j^p + w_i^q w_j^q + \ldots}{mn} \left(\frac{a^2}{2\pi(r_i^2 + r_j^2)}\right)^{3/2} \exp\left\{-\frac{a^2}{2\left(r_i^2 + r_j^2\right)}\left|x_i - y_j\right|^2\right\}. \quad (2)$$

Here, the following definitions apply:

$F(X,Y)$: overlap of property density functions of conformations *X* and *Y*
X,Y: matrices of spatial atom coordinates for the two molecules with dimension of $3 \times m$ and $3 \times n$, respectively
m, n: numbers of atoms in molecules *X* and *Y*, respectively
x_i: vector of coordinates of atom *i* in conformation *X*
w_i^P: weight of atom *i* with respect to property *p*: $w_i^P = 1$ if atom *i* has property *p*, otherwise $w_i^P = 0$
a: scaling factor; set to 2 in our calculations
r_i: van der Waals radius of atom *i*

Calculation of the atomic properties was performed with MOE. Atom properties were determined by pharmacophore types from MOE as described below.

$w_i^{aromatic} = 1$ if atom *i* is in a ring satisfying the Hueckel rule and is sp2 hybridized
$w_i^{donor} = 1$ if atom *i* is in pharmacophore class "Donor" or in class "Base"
$w_i^{acceptor} = 1$ if atom *i* is in pharmacophore class "Acceptor" or in class "Acid"
$w_i^{hydrophobic} = 1$ if atom *i* is in pharmacophore class "Hydrophobe"

A final normalization was carried out in order to obtain 3D similarity values between 0 (distinct spatial arrangement with no common atom positions) and 1 (identical conformation and position). The final 3D similarity values were obtained by dividing the overlap of the molecular property density functions by the mean self-overlap of the respective conformations:

$$F^{norm}(X,Y) = \frac{F(X,Y)}{\frac{1}{2}[F(X,X) + F(Y,Y)]}. \quad (3)$$

Supplemental Data

Supplemental Data include additional structural representations and a summary of the X-ray structures used in our analysis and are available at http://www.chembiol.com/cgi/content/full/14/5/489/DC1/.

Received: February 5, 2007
Revised: March 9, 2007
Accepted: March 14, 2007
Published: May 29, 2007

REFERENCES

1. Alaimo, P.J., Shogren-Knaak, M.A., and Shokat, K.M. (2001). Chemical genetic approaches for the elucidation of signaling pathways. Curr. Opin. Chem. Biol. *5*, 360–367.
2. Dobson, C.M. (2004). Chemical space and biology. Nature *432*, 824–828.
3. Butcher, R.A., and Schreiber, S.L. (2005). Using genome-wide transcription profiling to elucidate small-molecule mechanism. Curr. Opin. Chem. Biol. *9*, 25–30.
4. Klebe, G. (2000). Recent developments in structure-based drug design. J. Mol. Med. *78*, 269–281.
5. Eckert, H., and Bajorath, J. (2007). Molecular similarity analysis in virtual screening. Drug Discov. Today *12*, 225–233.
6. Kubinyi, H. (1998). Similarity and dissimilarity. A medicinal chemist's view. Perspect. Drug Discov. Des. *9–11*, 225–252.
7. Bajorath, J. (2002). Integration of virtual and high-throughput screening. Nat. Rev. Drug Discov. *1*, 882–894.
8. Schneider, G., Schneider, P., and Renner, G. (2006). Scaffold hopping. How far can we jump? QSAR Comb. Sci. *25*, 1162–1171.
9. Wang, R., Fang, X., Lu, Y., and Wang, S. (2004). The PDBbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures. J. Med. Chem. *47*, 2977–2980.

10. Wang, R., Fang, X., Lu, Y., Yang, C.-Y., and Wang, S. (2005). The PDBbind database: methodologies and updates. J. Med. Chem. *48*, 4111–4119.

11. Maggiora, G.M. (2006). On outliers and activity cliffs—why QSAR often disappoints. J. Chem. Inf. Model. *46*, 1535.

12. Boström, J., Hogner, A., and Schmitt, S. (2006). Do structurally similar molecules bind in a similar fashion? J. Med. Chem. *49*, 6716–6725.

13. Willett, P. (2005). Searching techniques for databases of two- and three-dimensional chemical structures. J. Med. Chem. *48*, 4183–4199.

14. Labute, P., Williams, C., Feher, M., Sourial, E., and Schmidt, J.M. (2001). Flexible alignment of small molecules. J. Med. Chem. *44*, 1483–1490.

15. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. (2000). The Protein Data Bank. Nucleic Acids Res. *28*, 235–242.

