Atom-Centered Interacting Fragments and Similarity Search Applications

José Batista,[†,‡] Lu Tan,[‡] and Jürgen Bajorath*[,‡]

In-Silico Center, JADO Technologies GmbH, Tatzberg 47-51, D-01307 Dresden, Germany, and 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

Received October 30, 2009

Parts of ligands in complex crystal structures that are involved in well-defined protein−ligand interactions are extracted and encoded as ensembles of atom-centered fragments, termed atom-centered interacting fragments (AIFs), which implicitly capture three-dimensional interaction information. AIF reference sets are utilized for feature count-based ranking of databases to search for molecules having similar activity. AIF calculations are reported for eight enzyme targets with multiple crystallographic inhibitor complexes and, in addition, for a complex of a G protein coupled receptor with an antagonist. The AIF approach further increases compound recall of structural key-based fingerprints. Moreover, AIF combinations are found to be specific markers of different classes of active compounds that lead to an early enrichment of inhibitors and selective antagonists in similarity search calculations.

INTRODUCTION

Similarity searching using two-dimensional (2D) molecular fingerprints is one of the traditional approaches to chemical database mining and continues to be a topic of considerable interest in chemoinformatics and drug discovery.[1] Original 2D fingerprint designs included atom pair[2] and structural fragment[3] fingerprints that are popular search tools to this date, for example, in the form of molecular access system (MACCS) structural keys.[4] The publicly available version of MACCS keys consists of a dictionary of 166 predefined molecular fragments. In addition to 2D fingerprints, three-dimensional (3D) pharmacophore fingerprints[5] have also been introduced for similarity searching. While 2D fingerprints are calculated from molecular graph representations, 3D fingerprints are derived from molecular conformations. Other types of 3D fingerprints have also been introduced that directly capture protein−ligand interaction information extracted from 3D complex structures.[6−8] Thus, conventional 2D and 3D fingerprints and interaction fingerprints are distinct in their design and represent 2D and 3D structural information in different ways.

Recently, attempts have been made to augment classical 2D fingerprint searching with 3D interaction information, without modifying the basic 2D fingerprint format.[9,10] Therefore, molecular fragments involved in well-defined protein−ligand interactions were extracted from crystallographic ligands, termed interacting fragments (IFs).[9] For IFs, MACCS keys were calculated to generate IF fingerprints (IF-FPs). In similarity search calculations, IF-FPs were found to perform equally well or better than MACCS fingerprints calculated for complete crystallographic ligands or other reference molecules.[9,10]

In contrast to interaction fingerprints, IF-FPs take 3D interaction information implicitly into account by transforming IFs into classical MACCS key representations. While MACCS keys have provided a convenient format for the evaluation of the IF approach, the dependence of IF encoding on structural key dictionaries (i.e., sets of predefined keys) also has potential shortcomings, in particular, taking into account that the IF of a ligand often consists of a set of disjoint fragments that might yield structural keys not originally present in the molecule.[9]

To generate a fragment dictionary-independent and unique encoding of interacting fragment information that retains the chemical information associated with individual atoms, we introduce herein the calculation of atom-centered fragments (AFs) for IFs, that is, atom-centered interacting fragments (AIFs). This approach yields IF-specific fragment ensembles, rather than a fixed-format fingerprint, which are found to be more suitable for the exploration of compound class-dependent structure−activity relationships than MACCS keys. We evaluate the AIF encoding on eight target enzymes with multiple crystallographic ligand complexes and, in addition, demonstrate its utility by carrying out another case study on a complex of a G protein coupled receptor (GPCR) with a selective antagonist.

MATERIALS AND METHODS

Generation of Interacting Fragments. From X-ray structures of target−ligand complexes, IFs were extracted using the molecular operating environment (MOE).[11] Ligand atoms involved in hydrogen bonding, ionic, and van der Waals interactions with cutoff distances of 3.2, 4.0, and 3.8, respectively, were isolated, and noninteracting atoms were omitted.[9] Figure 1 illustrates the generation of IFs, which can consist of disjoint fragment subsets.

* Corresponding author phone: +49-228-2699-306; fax: +49-228-2699-341; e-mail: bajorath@bit.uni-bonn.de.
† JADO Technologies GmbH.
‡ Rheinische Friedrich-Wilhelms-Universität.

10.1021/ci9004223
Published on Web 12/23/2009

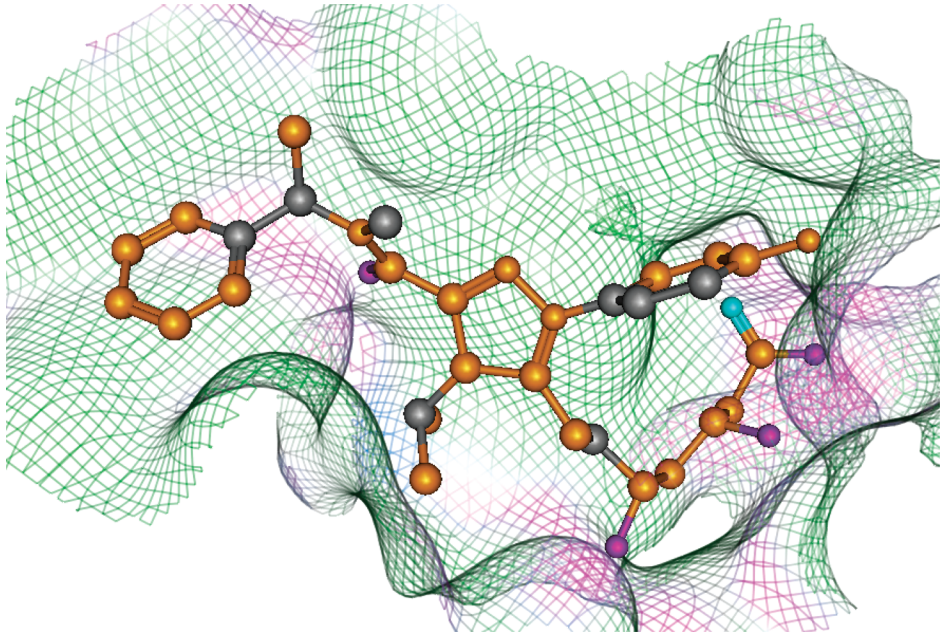

Figure 1. Interacting fragments. Interacting fragments are determined on the basis of hydrogen bond, van der Waals, and ionic interactions with predefined cutoff distances. Shown is an example for a HMG-CoA reductase−inhibitor complex (PDB id 2R4F). Green protein surface regions correspond to residues involved in van der Waals interactions and magenta regions to residues involved in hydrogen-bond interaction with the bound inhibitor. Ligand atoms involved in van der Waals interactions are colored gold, hydrogen-bond interactions are magenta, and ionic interactions are cyan. These atoms form the interacting fragment (IF). Non-IF atoms are shown in gray.

Atom-Centered Interacting Fragment Representation. Atom-centered fragments calculated herein are an extension of "augmented atoms" developed by Adamson et al.[12] that combine each atom in a molecule with its direct neighbors using the molecular graph as input. Here, each atom of a complete compound or an interacting fragment is once considered as a central atom, and atoms within bond distances of 2, 4, or 6 are added to this central atom. Hence, for any atom, three alternative AFs (compound) or AIFs (interacting fragment) are obtained. For each bond distance, the union of all AFs or AIFs yields the AF set for the complete ligand or AIF set for the interacting fragment, respectively. Atom and bond type information is retained. Exemplary fragments with bond distance 2 are depicted in Figure 2. In this figure, atoms forming interacting fragments are shown in bold (black). When calculating AFs/AIFs, the shortest path for each atom to the central atom is determined, and the atoms are organized in bond distance layers. Next, atoms within a given bond distance, for example, 4, corresponding to four layers, are combined. Because AFs are calculated for complete molecules, they might contain both noninteracting and interacting atoms. By contrast, AIFs exclusively consist of interacting atoms. If an AF or AIF is generated multiple times, it is collected only once. AFs and AIFs for each ligand are stored as SMARTS strings.[13] AIFs are collected only once because a library of unique descriptors is generated for binary encoding in a fingerprint format. Alternatively, count fingerprints could also be generated. However, we observed that only very few AIFs occurred multiple times. These AIFs were among the most generic ones.

Similarity Analysis. For a set of crystallographic ligands, IFs are determined and AIFs calculated (and also AFs for control calculations). Next, AIFs from individual ligands are pooled, forming the AIF reference set consisting of unique IFs. For compounds in a screening database, individual AF sets are calculated and compared to the reference set. Matching AIFs and AFs are counted to produce a database ranking in the order of decreasing numbers of matching fragments, following a similarity evaluation strategy previously established for feature ensemble fingerprints.[14] Because the number of generated AIFs is limited in many cases, standard fingerprint similarity metrics are difficult to apply. The AIF-based similarity search process is summarized in Figure 3.

Enzyme−Inhibitor Complexes. Eight target enzymes were selected for which multiple complex structures with noncovalent inhibitors were available in the Protein Data Bank (PDB).[15] For each enzyme, 10 unique complexes with highest resolution were selected, as summarized in Table 1. For each ligand, the IF was determined, and AIFs were calculated. In addition, we also studied a GPCR target, the human A_{2A} adenosine receptor in complex with a selective antagonist, termed ZM241385 (PDB code 3EML, resolution 2.6 Å).[16] For the bound antagonist, the AIF representation was also generated.

Compound Data Sets and Search Calculations. As a background database for similarity searching, 100 000 compounds were randomly selected from the ZINC database.[17] For each of the eight target enzymes, between 59 and 640 known inhibitors (Table 1) with pairwise MACCS Tanimoto similarity (Tc)[18] of ≤ 0.80 were selected from the MDL/Symyx Drug Data Report (MDDR)[19] and added to the screening database as potential hits.

For GPCR antagonist searching, different types of known GPCR antagonists were also selected from the MDDR (and added to the background database) including 89 antagonists selective for the adenosine A_{2A} receptor, 36 antagonists active against multiple adenosine receptor subtypes including A_{2A} (i.e., promiscuous adenosine antagonists), and 785 antagonists of other GPCRs with no reported activity against adenosine receptors.

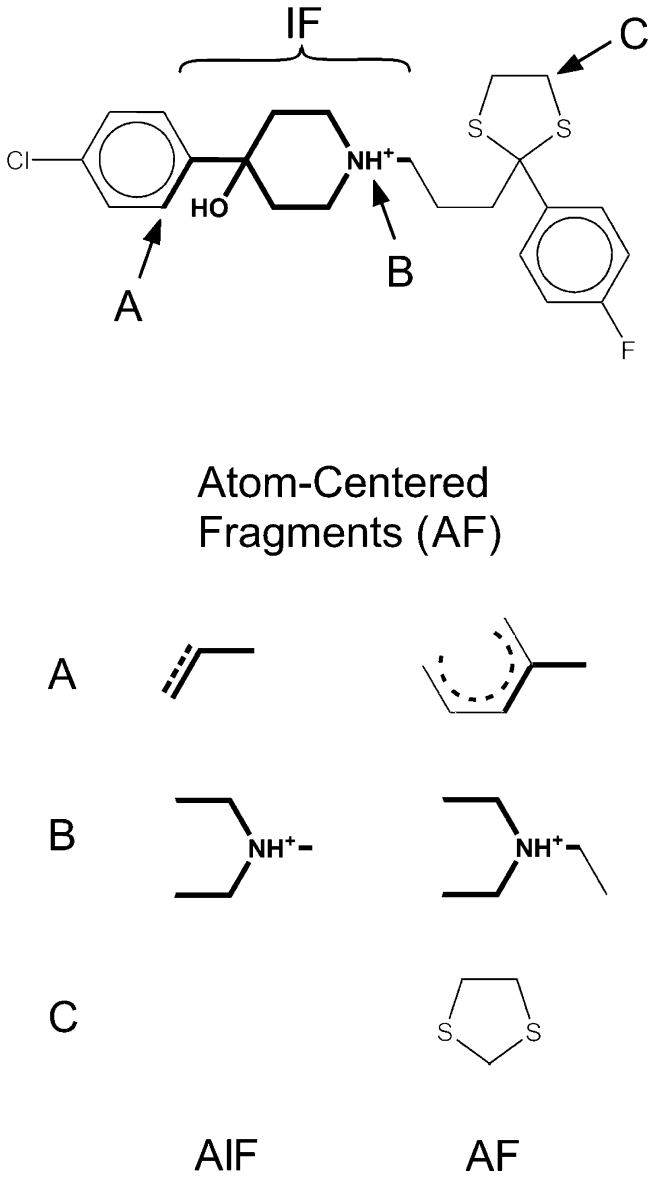

Figure 2. Atom-centered fragments. The generation of atom-centered fragments (AFs) and atom-centered interacting fragments (AIFs) for bond distance 2 is illustrated for a protease inhibitor. The atoms and bonds forming the IF are shown in bold. The AF/AIF generation is illustrated for three atoms labeled A, B, and C. For each of these atoms, the resulting fragments are composed of all atoms that are within a two-bond distance from the central atom. Dashed lines indicate aromatic bonds.

As control calculations for AIF feature counting, AF calculations were carried out for crystallographic ligands. Additional control calculations were also carried out with MACCS-based IF-FPs.[9] For these calculations, individual IF-FPs were generated for crystallographic reference molecules of each target enzyme and compared to MACCS fingerprints of database compounds in nearest neighbor (NN) searching[20] including 1-NN and 10-NN calculations. In 1-NN calculations, the highest *Tc* value generated by a reference molecule is utilized as the final similarity score of a database compounds, and in 10-NN calculations, the *Tc* values of all 10 reference compounds are averaged to produce the final similarity score.

For all search calculations, the recall of correctly identified active compounds within the 10 and 100 top-scoring database molecules was determined.

RESULTS AND DISCUSSION

Atom-Centered Interactive Fragments and Their Distribution. The interacting fragment approach was designed to implicitly incorporate 3D protein−ligand interaction information into ligand-based 2D similarity searching.[9] The encoding of IFs using structural key dictionaries provided a convenient format for IF representation and similarity search assessment. Because IFs are contained in the ligands they originate from, the observation that IF-FPs frequently improve similarity search performance as compared to structural key representations of entire ligands suggests that the reduction of ligands to interacting parts reduces the noise of 2D similarity search calculations. Consistent with this idea, IF-FPs are found to perform particularly well when multiple crystallographic ligands are utilized and common structural patterns emphasized through fingerprint scaling.[21]

An intrinsic drawback of a dictionary-based encoding of IFs is that interaction information is always transformed into predefined structural keys and that keys might be generated through disjoint IF components that do not appear in the original ligands. Furthermore, fragments that might be unique to a given compound class could not be generated. Therefore, with atom-centered fragments, we introduce an IF encoding that produces characteristic AIF sets for crystallographic ligands, as illustrated in Figure 2. These fragment sets can be utilized in similarity searching and are capable of producing signatures of different structure−activity relationships, as further discussed below.

In Table 2, the number of AFs and AIFs generated by each crystallographic reference set is reported. The number of AIFs per class for bond distance 2 ranged from 51 and 133 and was, as expected, generally smaller than the corresponding number of AFs. For some targets, the cumulative number of AIFs per set was much smaller, for example, HM (51 AIFs vs 133 AFs), TH (83 vs171), or XA (91 vs 151), hence indicating that parts of ligands were not involved in strong interactions, consistent with the general idea underlying the IF approach. For fragments with bond distance 4, the differences between the number of AIFs and AFs were overall even larger, for example, HM (35 vs 206) or TH (74 vs 216).

Similarity Searching. To test AIF representations in similarity searching, we implemented a simple feature count-based search strategy, as outlined in Figure 3. The results of our search and control calculations are summarized in Table 2 for atom-centered fragments with bond distances 2 and 4 (fragments with bond distance 6 displayed overall low search performance, and the results are therefore not reported). A few ligands that consistently failed to recover any active compounds were omitted from data analysis. These ligands typically contained core structures that differed from those in other active compounds and hence produced many unique AIFs.

First, we compared AIF representations with AFs for complete ligands. For fragments with bond distance 2, AIF compound recall was equal to or higher than AF recall for six of eight targets and lower in two cases. For bond distance 4, AIF recall was equal to or higher than AF recall for seven of eight classes. Overall, compound recall was higher for fragments with bond distance 4 than fragments with bond

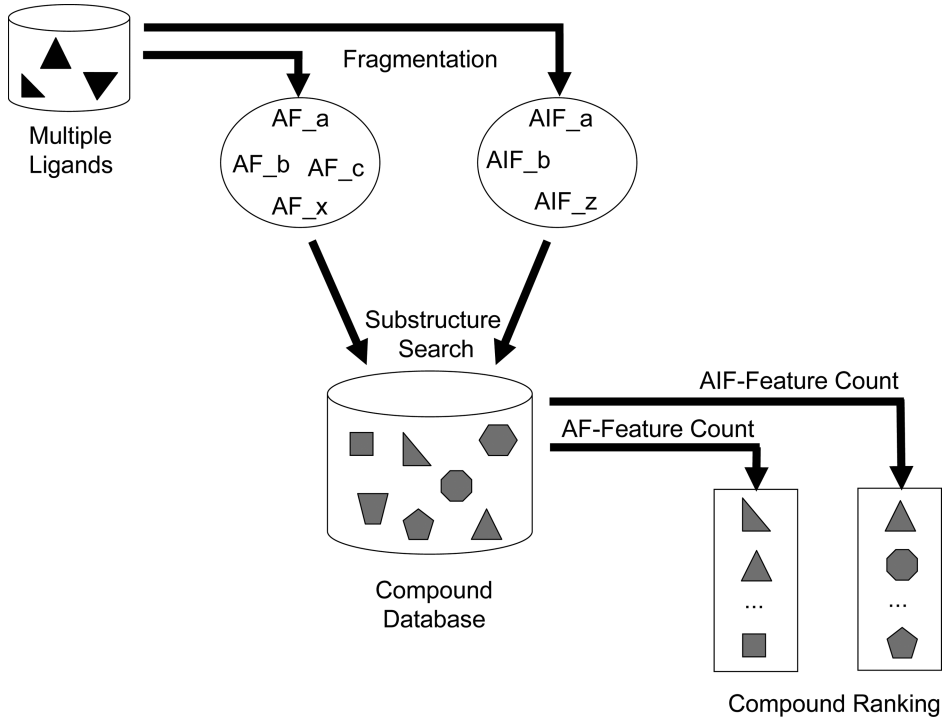

Figure 3. Similarity search strategy for atom-centered interacting fragment sets. Sets of AFs and AIFs are generated from multiple crystallographic ligands and utilized via a substructure search procedure for "feature count" ranking of database compounds. The more AFs or AIFs match a database compound, the higher the compound is ranked.

Table 1. Target Enzymes and Inhibitor Complex Structures[a]

target	no. MDDR	PDB id crystallographic resolution (Å)				
DP dipeptidyl amino peptidase IV	135	3C43	2QTB	2OPH	2FJP	2QT9
		2.30	2.25	2.40	2.40	2.10
		3C45	1N1M	3D4L	2P8S	2BUB
		2.05	2.50	2.00	2.20	2.66
DR dihydrofolate reductase	59	1DG5	1DAJ	1S3V	1AO8	1RF7
		2.00	2.30	1.80	1.70	1.80
		1CD2	3CSE	2C2S	1KLK	1J3I
		2.20	1.60	1.40	2.30	2.33
F7 factor VIIa	119	1WQV	1WSS	1WTG	1WV7	1YGC
		2.50	2.60	2.20	2.70	2.00
		1Z6J	2AEI	2C4F	2EC9	2ZP0
		2.00	2.52	1.72	2.00	2.70
HM HMG-CoA reductase	246	2R4F	2Q6C	3CDB	3CD7	3CD0
		1.70	2.00	2.30	2.05	2.40
		3CCZ	3CCW	3CDA	3CCT	3BGL
		1.70	2.10	2.07	2.12	2.22
P4 phospho-diesterase IV	640	1MKD	1XLX	1XLZ	1XM4	1XM6
		2.90	2.19	2.06	2.31	1.92
		1XMU	1XMY	1XN0	1XOQ	2FM0
		2.30	2.40	2.31	1.83	2.00
PR HIV-1 protease	234	1EC1	1EBY	1D4H	1DMP	1BDQ
		2.10	2.29	1.81	2.00	2.50
		1HBV	1EC0	1PRO	2AID	1NPV
		2.30	1.79	1.80	1.90	2.00
TH thrombin	238	1A4W	1BMM	1C1U	1C4V	1K21
		1.80	2.60	1.75	2.60	1.86
		1KTS	1OYT	1QUR	1VZQ	1W7G
		2.40	1.67	2.00	1.54	1.65
XA factor Xa	393	1EZQ	1F0S	1FAX	1FJS	1KSN
		2.20	2.10	3.00	1.92	2.10
		1KYE	1NFY	1X7A	1Z6E	2FZZ
		2.22	2.10	2.90	1.80	2.20

[a] Reported are the eight target enzymes and inhibitor sets used in this study and the number of MDDR compounds (no. MDDR) utilized as potential hits for similarity searching. Resolution is reported in angstroms, and PDB id gives the structure entry code for the Protein Data Bank.

Table 2. Recall of Active Compounds[a]

	IF-FP (1-NN search)		IF-FP (10-NN search)	
	top 10	100	top 10	100
DP	3	11	4	21
DR	4	13	1	10
F7	3	7	4	19
HM	0	2	8	31
P4	1	15	1	9
PR	0	1	1	3
TH	4	11	2	20
XA	2	9	5	29

	AF2			AF4		
	no. AFs	top 10	100	no. AFs	top 10	100
DP	117	3	18	134	9	22
DR	82	6	24	76	5	23
F7	159	1	7	218	3	10
HM	133	4	6	206	4	5
P4	95	9	86	120	8	86
PR	119	4	32	164	8	36
TH	171	5	30	216	6	35
XA	151	8	27	154	7	40

	AIF2			AIF4		
	no. AIFs	top 10	100	no. AIFs	top 10	100
DP	88	8	26	94	9	28
DR	81	7	17	61	5	25
F7	133	2	6	173	3	10
HM	51	10	62	35	9	69
P4	76	9	60	86	10	85
PR	100	8	44	111	8	43
TH	83	3	13	74	4	11
XA	91	7	33	83	9	57

[a] For similarity searching, the number of retrieved active compounds among the top 10 or 100 database compounds is reported for different molecular representations including results of 1- and 10-NN reference calculations using IF-FPs, atom-centered fragments (AF*n*), and atom-centered interacting fragments (AIF*n*). For comparison, fragments with bond distance $n = 2$ and $n = 4$ are shown. The number of AFs or AIFs in each reference set is also given.

distance 2 (see below). We also compared AIF compound recall with nearest neighbor search calculations using the original structural key-based IF-FPs (Table 2). Here, a consistent increase of AIF recall (bond distance 4) relative to IF-FP 1-NN calculations was observed and an increase over 10-NN calculations in six of eight cases. Overall, the recall achieved in AIF calculations was promising. In database selection sets of 100 molecules, between 6 and 62 active compounds were found for fragments with bond distance 2 (on average 32.6 compounds per target), and for fragments with bond distance 4, between 10 and 85 active compounds (on average 41.0 compounds).

In two instances, F7 and TH, AIFs did not improve AF search performance and performed slightly worse, respectively. However, in the case of F7, fewer AIFs than AFs produced essentially the same recall. For TH, many interacting fragments were found to consist of multiple small substructures. Consequently, the resulting AIFs were often more generic in nature than corresponding AFs and thus matched many database compounds, which then reached higher ranks.

Thus, taken together, systematic search calculations over eight different sets of crystallographic ligands revealed the tendency that AIF representations (interacting fragments) yielded higher recall of active database compounds than corresponding AFs (complete ligands) and IF-FPs (interacting fragments), although there were ligand set-dependent differences in relative search performance. Furthermore, AIF representations produced in part significant compound recall in cases where IF-FPs failed, or nearly failed, to recover active compounds, that is, HM, P4, or PR.

A characteristic feature of AIF calculations was a notable enrichment of active compounds in small selection sets of 10 database molecules, in marked contrast to IF-FPs. For example, for AIFs with bond distance 2, 8−10 active compounds were found within the 10 top-scoring database molecules in four cases (on average 6.8 active compounds), and for AIFs with bond distance 4 in five cases (on average 7.1 active compounds). A comparable enrichment was observed for AF representations with bond distance 4 (but not 2) where in three cases, eight or nine active compounds occurred among the top 10 database molecules (on average 6.2 compounds), although AIF enrichment was higher. Thus, in contrast to dictionary-based encoding, atom-centered fragments produced specific signatures of different structure−activity relationships that clearly distinguished subset of active compounds from background database molecules, especially for interacting fragments. This signature character of AIF encoding is further described in the following.

GPCR X-ray and Ligand Systems. The systematic search calculations over the eight crystallographic ligand sets of different targets permitted an overall assessment of the AIF. In addition, we also carried out a case study on an individual target−ligand complex that is currently of particular interest in pharmaceutical research: the X-ray structure of the human A_{2A} adenosine receptor in complex with the A_{2A} selective antagonist ZM241385 (ZM),[16] as shown in Figure 4. This structure is one of only four X-ray structures of GPCRs in complex with inverse agonists or antagonists that are currently available.[22] Different from β adrenergic receptors, which form the other three complexes, we have been able to assemble a ligand reference system for adenosine receptors consisting of A_{2A}-selective antagonists, promiscuous adenosine receptor antagonists, and antagonists active against other GPCRs (see Materials and Methods).

Considering its structural details, the A_{2A}−ZM complex has been an interesting test case for the AIF approach. As shown in Figure 4, in the X-ray structure, which represents an average of an extensive conformational ensemble available to GPCRs,[22] the bound antagonist consists of a strongly interacting part (encompassing the purine and furan rings) and a very weakly interacting part (the methyl phenol moiety), yielding a well-defined interacting fragment, as shown in Figure 4. Hence, a key question has been how the absence of the methyl phenol moiety of ZM might affect search calculations using this individual query.

Adenosine Receptor Antagonist Searching. To investigate this question, we generated AF representations for ZM and AIF representations for the interacting fragment shown in Figure 4. As reported in Table 3, AFs with bond distance 2 and 4 produced 22 and 20 fragments, respectively, whereas AIFs with bond distance 2 and 4 generated 15 and 12 fragments, respectively. On the basis of these

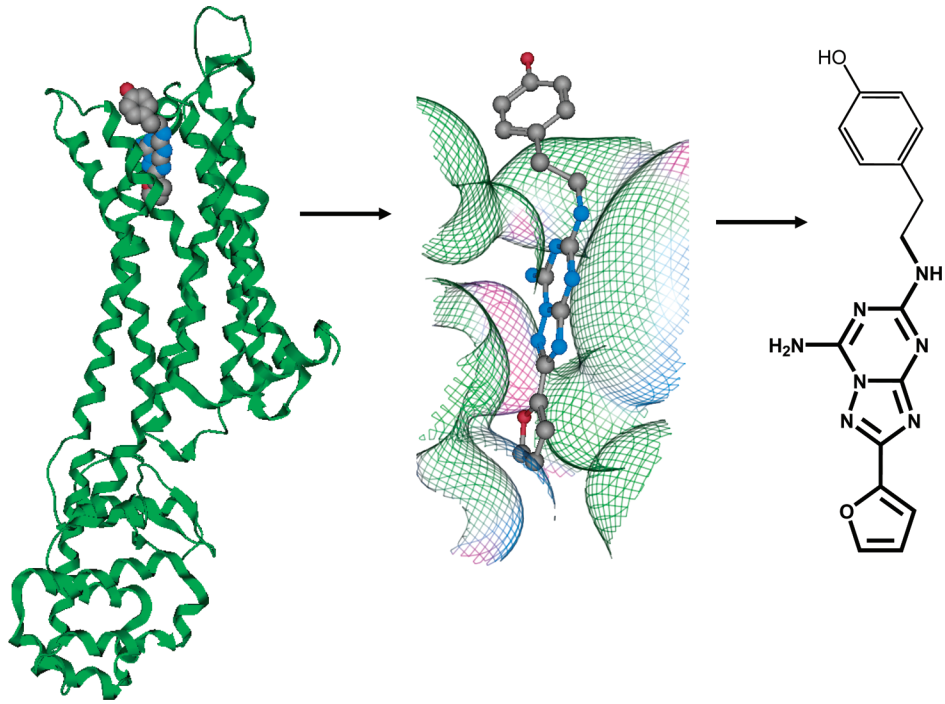

Figure 4. Adenosine A_{2A} receptor−antagonist complex. On the left, the X-ray structure of the adenosine receptor with bound antagonist ZM241385 is shown, and in the middle is a close-up view of the ligand binding site. On the right, the 2D structure of the antagonist is displayed with its IF shown in bold.

Table 3. Feature Ranking of GPCR Antagonists and ZINC Compounds[a]

encoding	no. features	feature counts	selective	promiscuous	other GPCR ligands	ZINC
AIF2	15	15	4	2	0	0
		10	3	1	0	0
		9	0	0	0	1
		8	3	1	0	0
		7	6	0	3	0
		6	5	0	0	6
		total	**21**	**4**	**3**	**7**
		5	3	1	2	40
AF2	22	18	0	2	0	0
		14	5	0	0	0
		12	0	1	0	0
		11	1	1	0	4
		10	2	0	2	12
		total	**8**	**4**	**2**	**16**
		9	1	0	2	74
AIF4	12	12	5	1	0	0
		4	3	1	0	1
		3	1	2	1	1
		total	**9**	**4**	**1**	**2**
		2	12	4	2	151
AF4	20	15	2	0	0	0
		11	3	0	0	0
		9	1	0	0	0
		6	0	0	0	7
		total	**6**	**0**	**0**	**7**
		5	1	1	46	862

[a] For AIFs and AFs with bond distance 2 and 4, database compounds are ranked in the order of decreasing features, and the ranking is reported up to the feature level where a notable increase in the number of matching ZINC decoys is detected. The row "total" reports the number of different compounds ranked above this level.

features sets, the background database containing the GPCR antagonist collection was ranked. In Table 3, we report the number of fragments that were matched by top-ranked compounds and the type of these compounds. In each case, a number of selective and promiscuous GPCR antagonists were detected prior to ZINC compounds. For bond distance 2, the complete set of 15 AIFs was matched by six database compounds including four selective A_{2A} and two promiscuous adenosine receptor antagonists. The next database compounds matched only 10 of 15 fragments including three selective and one promiscuous antagonists. For the corresponding AF representations, two promiscuous antagonists matched 18 of 22 fragments and five selective antagonists 14 of 22 fragments. Similar observa-

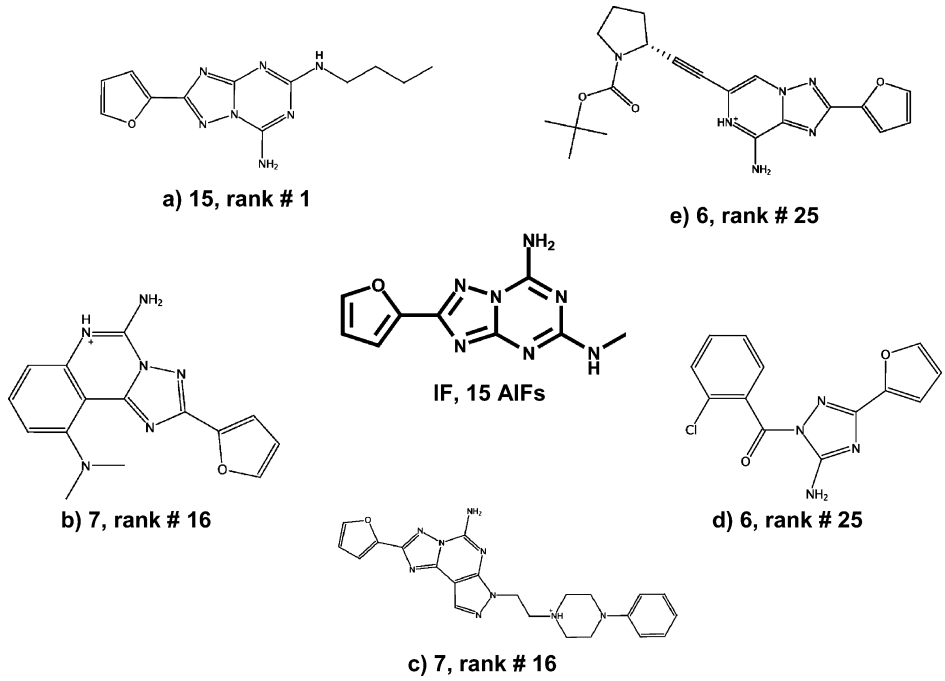

Figure 5. Diverse antagonists. The IF of ZM241385 is shown in the center, producing 15 reference AIFs (for bond distance 2). Compounds a−e are structurally diverse selective A_{2A} antagonists detected in the similarity search. Reported is the number of reference AIFs each molecule matches, followed by its rank.

tions were made for AIFs/AFs with bond distance 4. Thus, the feature matching procedure was highly selective and mirrored the early enrichment characteristics detected in the systematic search trials described above. This means that these fragment combinations also produced signatures of antagonist structure−activity relationships. Moreover, as illustrated in Figure 5, highly ranked antagonists contained different scaffolds, demonstrating that the AIF

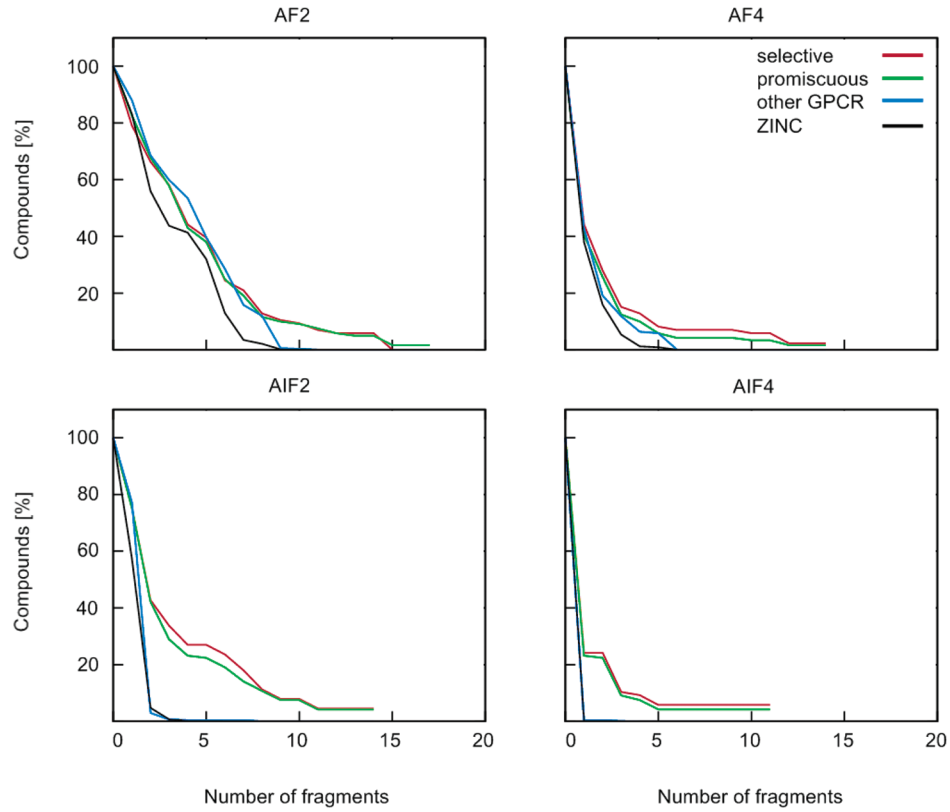

Figure 6. Fragment distribution. The AF and AIF distributions in selective A2A, promiscuous adenosine, and other GPCR antagonists and ZINC decoys are shown. For example, AF2 shows the distribution of AFs with bond distance 2 (see Table 3). On the *x*-axis, the number of fragments present in individual compounds is reported, and on the *y*-axis, the percentage of compounds in each set containing these fragments is shown. Distributions are represented in a cumulative manner; that is, compounds containing 10 fragments are also included in those that contain nine or fewer.

matching procedure was capable of enriching structurally diverse antagonists.

The compound rankings reported in Table 3 also show that AIFs produced higher recall of relevant antagonists than AFs (in this case, AIF2 encoding was superior to AIF4). Thus, the interacting fragment indeed provided a more specific query than the complete antagonist. This is also evident in the fragment distributions shown in Figure 6, which also reflect the signature character of fragment combinations. For the AIF encoding with bond length 2, the percentage of ZINC decoys and other GPCR antagonists containing four or more fragments approached 0 when approximately 27% of selective or promiscuous antagonists were retrieved. By contrast, for the corresponding AF encoding, the percentage of ZINC decoys or other GPCR antagonists matching 10 or more fragments approached 0, but at this stage, only approximately 10% of selective or promiscuous antagonists were detected. Thus, although both AIF and AF encodings yielded a notable enrichment of selective and promiscuous antagonists among top-ranked database compounds, the AIF representation produced higher recall of desirable antagonists.

CONCLUDING REMARKS

In this study, we have transformed interacting fragments of active compounds extracted from X-ray structures into sets of atom-centered interacting ligand fragments, which provide an independent encoding of interacting fragment information. AIF reference sets were explored in systematic similarity search calculations applying a simple feature count strategy for database ranking. In test calculations on different sets of enzyme inhibitors, AIF representations were effective search tools and displayed signature character leading to early enrichment of inhibitors in database selection sets of small size. Interacting fragment information augments 2D similarity searching, and the molecule-specific AIF encoding, which yields variable fragment sets, was found to be superior to structural key-based interacting fragment fingerprints. Furthermore, the A_{2A} adenosine receptor complex with ZM has provided an intuitive example for the utility of the AIF approach and has been explored at the molecular level of detail. In this case, the preferential and specific recognition of desirable antagonists over other database compounds was also observed, consistent with the signature character of fragment combinations. In future studies, the concept of implicitly adding 3D interaction information to ligand-based similarity searching might be further extended, for example, by weighting different protein−ligand interactions, and other encodings of interacting fragment information might also be explored. The molecular benchmark systems designed for our study are freely available upon publication via the following URL: http://www.lifescienceinformatics.uni-bonn.de.

ACKNOWLEDGMENT

L.T. is supported by a fellowship of the Graduiertenkolleg (GRK) 804 of the Deutsche Forschungsgemeinschaft.

REFERENCES AND NOTES

(1) Willet, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
(2) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64–73.
(3) Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest neighbor searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36–41.
(4) *MACCS Structural Keys*; Symyx Software: San Ramon, CA, 2005.
(5) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview over the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
(6) Deng, Z.; Chuaqui, C.; Singh, J. Structural interaction fingerprint (SIFt): a novel method for analyzing three-dimensional protein-ligand binding interactions. *J. Med. Chem.* **2004**, *47*, 337–344.
(7) Kelly, M. D.; Mancera, R. L. Expanded interaction fingerprint method for analyzing ligand binding modes in docking and structure-based drug design. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1942–1951.
(8) Baroni, M.; Cruciani, G.; Sciabola, S.; Perruccio, F.; Mason, J. S. A common reference framework for analyzing/comparing proteins and ligands. Fingerprints for ligands and proteins (FLAP): theory and application. *J. Chem. Inf. Model.* **2007**, *47*, 279–294.
(9) Tan, L.; Lounkine, E.; Bajorath, J. Similarity searching using fingerprints of molecular fragments involved in protein-ligand interactions. *J. Chem. Inf. Model.* **2008**, *48*, 2308–2312.
(10) Tan, L.; Bajorath, J. Utilizing target-ligand interaction information in fingerprint searching for ligands of related targets. *Chem. Biol. Drug Des.* **2009**, *74*, 25–32.
(11) *Molecular Operating Environment (MOE), Version 2008. 10*; Chemical Computing Group, Inc.: Montreal, Quebec, Canada, 2008.
(12) Adamson, G. W.; Lynch, M. F.; Town, W. G. Analysis of structural characteristics of chemical compounds in a large computer-based file. Part II. Atom-centered fragments. *J. Chem. Soc. C* **1971**, 3702–3706.
(13) *SMARTS*; Daylight Chemical Information Systems, Inc.: Aliso Viejo, CA, 2008.
(14) Hu, Y.; Lounkine, E.; Bajorath, J. Filtering and counting of extended connectivity fingerprint features maximizes compound recall and the structural diversity of hits. *Chem. Biol. Drug Des.* **2009**, *74*, 92–98.
(15) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
(16) Jaakola, V. P.; Griffith, M. T.; Hanson, M. A.; Cherezov, V.; Chien, E. Y.; Lane, J. R.; Ijzerman, A. P.; Stevens, R. C. The 2.6 angstrom crystal structure of a human A2A adenosine receptor bound to an antagonist. *Science* **2008**, *322*, 1211–1217.
(17) Irwin, J. J.; Shoichet, B. K. ZINC-a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(18) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.
(19) *Molecular Drug Data Report (MDDR), version 2005. 2*; Symyx Technologies, Inc.: Sunnyvale, CA, 2005.
(20) Hert, J.; Willett, P.; Wilton, D. J. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(21) Tan, L.; Vogt, M.; Bajorath, J. Three-dimensional protein-ligand interaction scaling of two-dimensional fingerprints. *Chem. Biol. Drug Des.* **2009**, *74*, 449–456.
(22) Rosenbaum, D. M.; Rasmussen, S. G. F.; Kobilka, B. K. The structure and function of G-protein-coupled receptors. *Nature* **2009**, *459*, 356–363.

CI9004223