Distinguishing between Bioactive and Modeled Compound Conformations through Mining of Emerging Chemical Patterns

Jens Auer and Jürgen Bajorath*

Department of Life Science informatics, B-IT, LIMES Program Unit Chemical Biology & Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstrasse 2, D-53113 Bonn, Germany

Received May 27, 2008

To systematically compare bioactive and theoretically derived compound conformations, we have analyzed 18 different sets of active small molecules with experimentally determined binding conformations and modeled conformers using a pattern recognition approach. Compound class-specific descriptor value range patterns that accurately distinguish bioactive conformations from other low-energy conformers were identified for all 18 compound classes. Discriminatory patterns were often chemically intuitive and could be well rationalized on the basis of X-ray structures of the protein−ligand complexes. Target-specific descriptor patterns can be used as filters to screen conformational ensembles for bioactive conformations.

INTRODUCTION

Prediction of the bioactive conformations of small molecules continues to be a major challenge in molecular design, QSAR modeling, or pharmacophore analysis. Similarly, distinguishing between binding conformations of active compounds and other low-energy conformers represents a comparably difficult task. Over the past decade, several studies have investigated binding conformations of known active compounds, mostly enzyme inhibitors, taken from complex crystal structures.[1−6] A major focal point of these investigations has been the analysis of intramolecular strain energy of small molecules that is generally induced upon protein binding. It is well appreciated that ligands do not bind in global energy minimum conformations to their targets because achieving a high degree of molecular complementarity within a binding or active site generally comes at the cost of steric strain. The strain energy penalty associated with the formation of protein−ligand complexes can be approximated by computational means. For example, depending on the force field used to calculate relevant energy terms, Perola and Charifson have estimated total strain energy of average small molecular ligands to be approximately 2 kcal/mol.[2] Steric strain effects contribute to the difficulties associated with correctly predicting bioactive ligand conformations, which is often attempted by systematic conformational sampling and filtering of low-energy conformers. It is therefore not surprising that strain energy and its consequences have been intensely studied.[1−6] Going beyond the analysis of strain effects, only very few studies have attempted to systematically explore differences between binding and modeled conformations. For example, in a pioneering study reported in 2002, Diller and Merz[1] compared 65 small molecules taken from X-ray structures of protein−ligand complexes to 5000 low-energy conformations. For each experimental and corresponding energy-minimized conformation, the distribution of the values of six three-dimensional (3D) descriptors was compared. These descriptors included polar and apolar solvent-accessible surface area, the radius of gyration, dipole moment, the number of intermolecular interactions, and the ratio of two principal molecular axes. It was found that binding conformations tended to have larger solvent-accessible surface area than minimized conformations because of fewer intramolecular interactions. Binding conformations were in general also found to be less compact than energy-minimized ones.

One would hope that systematic comparisons of bioactive and modeled ligand conformations for different targets might ultimately help to identify active conformations in conformational ensembles, which is, as mentioned above, of paramount importance for various ligand-based drug design strategies. This would require deducing target-specific rules or feature combinations that could differentiate between alternative conformations.

To address such issues, we have followed up on the theme of the analysis by Diller and Merz and evaluated an in-house developed pattern recognition and data mining approach for its potential to identify compound class-specific descriptor value range patterns (i.e., signature patterns) that distinguish bioactive conformations from other low-energy conformers with high accuracy. We have studied sets of inhibitors for 18 different target proteins and, in each case, have been able to identify patterns that correctly identify bioactive conformations and differentiate them from others, even if conformational differences were subtle. Furthermore, it has been possible to rationalize key patterns at the molecular level of detail by analysis of X-ray structures of enzyme−inhibitor complexes.

METHODS

Test Compounds in Bioactive Conformations. Inhibitors available in complex crystal structures with 18 proteins (including the FK506 binding protein and 17 enzymes) were analyzed in our study, as summarized in Table 1. For each target, multiple inhibitors in their X-ray binding conforma-

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

10.1021/ci8001793 CCC: $40.75
Published on Web 08/13/2008

Table 1. Inhibitors, Modeled Conformers, and Discriminatory Patterns[a]

class	cmpds	modeled conformers	rmsd range	rmsd average	patterns
acetylcholine esterase	5	16	0.69−5.39	2.69	15
adenosine deaminase	15	139	1.10−4.18	2.45	17
carbonic anhydrase	21	198	0.57−4.30	2.26	36
carboxypeptidase	8	60	0.73−5.03	2.71	17
cyclin dep. kinase	31	247	0.70−4.08	2.24	7
elastase	3	30	1.59−4.09	2.69	81
endothiapepsin	6	19	3.30−5.87	4.80	33
factor Xa	10	91	0.94−5.68	2.52	32
FK506 BP	6	51	1.82−4.23	3.06	15
HIV protease	20	142	0.98−6.27	3.69	8
plasminogen activator	7	43	0.25−2.76	1.69	8
PT phosphatase1b	14	67	0.30−4.58	2.40	17
protocat. -dioxygenase	10	16	0.08−3.39	0.87	10
ribonuclease A	9	45	0.98−3.39	2.12	10
thermolysin	6	42	2.05−4.95	3.30	197
thrombin	21	158	0.60−4.51	2.69	7
trypsin	30	195	0.33−3.37	1.74	5
tyrosine kinase	5	23	2.72−5.47	3.95	30

[a] Reported are the number of inhibitors with experimental binding conformations per class (cmpds) and the number of modeled low-energy conformers, their rmsd range relative to the corresponding experimental conformation, the average rmsd per class, and the number of emerging chemical patterns (patterns) that discriminate between bioactive and modeled conformations. Cyclin-dep. kinase stands for cyclin-dependent kinase; FK506 BP stands for FK506 binding protein, and protocat.-dioxygenase stands for protocatechuate-3,4-dioxygenase.

tions were taken from the PDBBind database.[7,8] A total of 227 active compounds were collected, ranging from 3 to 30 inhibitors per set (Table 1). These 18 different sets of compounds were used for pattern analysis.

Modeled Conformations. Each active compound was subjected to extensive conformational search using the Molecular Operating Environment (MOE, version 2007.09).[9] A stochastic conformational search was carried out for 10 000 iterations by randomly rotating single bonds in test molecules. Following each iteration, the resulting conformation was energy minimized and sampled. Cartesian minimization was carried out using MOE's MMFF94x force field until the rms gradient of the energy function was less than 0.001 kcal mol^{-1} $Å^{-1}$. For each inhibitor, sampled low-energy conformations were compared to eliminate conformations from pairs of very similar ones, applying a root-mean-square deviation (rmsd) threshold value of 0.1 Å. As reported in Table 1, between 16 and 247 low-energy conformations were retained per class. The rmsd values for modeled and experimental conformations were calculated on the basis of superposition of all non-hydrogen atoms.

Descriptors. A total of 67 conformation-dependent (3D) descriptors available in MOE were used in our study. Supporting Information Table 1 specifies this descriptor set that includes a variety of 3D descriptors belonging to five categories: energy, shape, and charge distribution descriptors, molecular surface properties, and volume-dependent descriptors. A subset of these descriptors including, for example, heat of formation, ionization potential, and highest-occupied molecular orbital (HOMO) or lowest-unoccupied molecular orbital (LUMO) energies was calculated through MOE's interface with MOPAC[10] and the semiempirical AM1,[11] PM3,[12] and MNDO[13] methods. Partial charge and potential energy descriptors were calculated with the MMFF94x force field. The radius of gyration (rgyr) serves as a measure of compactness of a molecule and the principal moments of inertia (pmi, pmiX, pmiY, pmiZ) account for mass distribution. Surface descriptors include, for example, the total solvent-accessible surface area (ASA) and positively or negatively charged (ASA+ and ASA−), hydrophobic (ASA_H) or polar (ASA_P) surface areas. In addition, descriptors depending on the van der Waals volume (vol) are also included, such as molecular density (dens) that is calculated by dividing molecular weight by vol. Because the calculation of some of these descriptors depends on Cartesian rather than internal coordinates, all test compounds were translated into a reference coordinate system prior to descriptor calculation, with their center of mass matching the origin. A standard orientation of test molecules was obtained by aligning the first principal molecular axis with the *x*-axis, the second with the *y*-axis, and the third with the *z*-axis of the reference coordinate system. The calculation of some of these 3D descriptors, such as MOPAC descriptors, is computationally expensive. However, all descriptors need to be calculated only once, which does not limit the applicability of the method.

Discretization. For our pattern mining analysis, as described below, continuous descriptor value ranges must be transformed into discrete subranges. For this purpose, we have applied a technique that divides a descriptor value range into discrete intervals on the basis of the information entropy of the ensuing data distribution in these intervals,[14] as described previously.[15] Details of the algorithm are provided as Supporting Information. The discretization algorithm was applied to each individual set of bioactive conformations and low-energy conformers, as well as the combinations of all bioactive and modeled sets, respectively. The resulting descriptor patterns enabled both global and target-specific comparisons of experimental and modeled ligand conformations.

Mining of Descriptor Patterns. The pattern mining approach applied here aims at the identification of descriptor value range combinations that frequently occur in bioactive compound conformations but rarely in modeled ones. To systematically analyze and compare descriptor value ranges and derive patterns, we have previously adapted a data mining technique called emerging patterns[16] for chemoinformatics applications.[15,17] The resulting approach termed Emerging Chemical Patterns (ECP) employs combinations of set-specific discrete attribute-value pairs. Details of the methodology are provided as Supporting Information including Supporting Information Tables 2 and 3. Briefly, the algorithm searches for combinations of features that often occur in a specific compound set, but not a reference set. A *pattern* is defined as any combination of single or multiple descriptor value ranges, for example, {MW = [0.00, 500.00)} or {MW = [0.00, 500.00), HB-acc = [0.00, 10.00)}. The latter pattern means that a combination of two descriptor value ranges discriminates against the reference set, that is, molecular weight from 0 to 500 together with 0 to 10 hydrogen bond acceptors. How frequently a pattern occurs

in a data set is measured by its *support*, that is, the percentage of compounds matching the pattern. A pattern is considered class-specific if the fraction of support in both compound sets, the so-called *growth* or *growth rate*, is larger than a predefined threshold. If the pattern exclusively occurs in one of two compound sets, the growth is *infinite*. Most interesting are the patterns with smallest cardinality, because they are contained in all pattern supersets and are therefore most general (see Supporting Information).

Here we have applied a growth rate threshold of 50% to bioactive conformations, meaning that only patterns matching at least half of the binding conformations were calculated. For modeled conformations, a threshold of 10% was applied. These parameter settings ensured that each detected pattern was at least five times more frequent in bioactive than modeled conformations. For each inhibitor set, all patterns with a growth rate of at least 10 were analyzed.

Protein−Ligand Interactions. For the interpretation of key patterns, details of protein−ligand interactions in the X-ray structures of their complexes were studied and represented with the aid of two-dimensional (2D) interaction diagrams[18] calculated with MOE. For comparison, modeled low-energy conformers were superposed on experimental conformations of each inhibitor, and interaction diagrams were also analyzed for these hypothetical complexes.

RESULTS AND DISCUSSION

Differences between Modeled and Experimental Conformations. To model conformations of known inhibitors and generate conformational ensembles for comparison with bioactive conformations, a standard conformational search protocol was applied. Table 1 lists the rmsd values obtained for pairwise comparisons of modeled conformers and binding conformations. In 12 of 16 cases, the conformational ensembles contained conformers that were very similar to binding conformations, that is, within 1 Å rmsd. Thus, in many cases, differences between experimental and modeled conformations were rather subtle, which can also be appreciated in Figure 1. However, all ensembles also contained conformations that deviated from binding conformations by several Å rmsd. Most classes produced an average rmsd of around 2 Å, suggesting that modeled conformers were overall not dramatically different from binding conformations. Thus, we considered the sampled conformations suitable for our analysis because they represented a spectrum of conformers with varying degrees of similarity to the binding conformations.

Global Comparison of Binding and Modeled Conformations. A major goal of our analysis has been to search for compound class- or target-specific descriptor patterns to distinguish between bioactive and modeled conformations. As an initial test, prior to analyzing descriptor patterns for individual classes, we determined whether the ECP methodology was capable of generally differentiating between binding and modeled conformations. ECP was originally developed to classify 2D representations of active and inactive compounds[15] and was subsequently used to simulate sequential screening experiments.[17] The evaluation of 3D patterns represented a novel application.

Therefore, we first generated a set of 227 binding conformations by combining all 18 compound classes and a reference set containing all 1598 low-energy conformers and

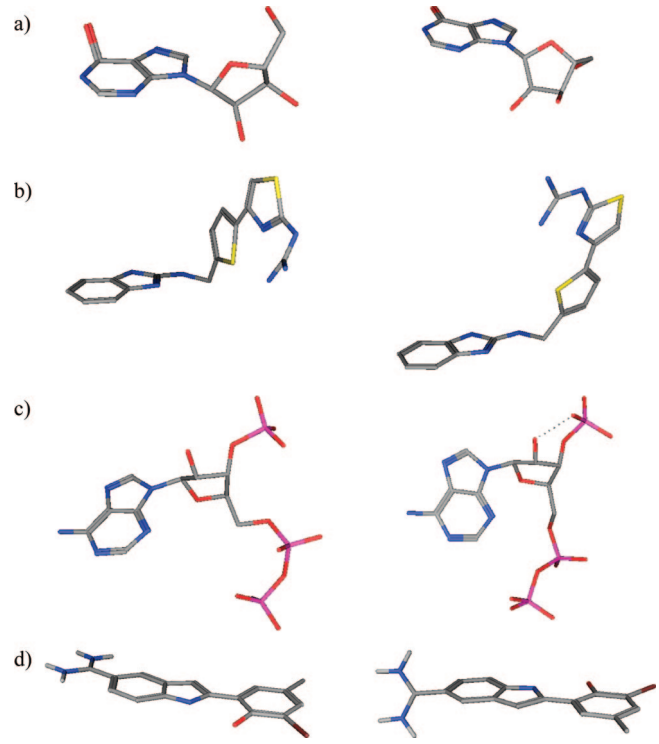

Figure 1. Examplary bioactive and modeled conformations. The binding conformations (on the left) and corresponding low-energy conformers (right) are shown for four inhibitors of three enzymes discussed in the text. The dashed line represents an intramolecular hydrogen bond: (a and b) adenosine deaminase inhibitors (taken from PDBind entries 1fkx and 1ndv, respectively), (c) ribonuclease A inhibitor (1afk), and (d) trypsin inhibitor (1o3h).

Table 2. Emerging Chemical Patterns for Discriminating the Combination of All Bioactive Conformations From All Modeled Conformations[a]

growth	*B* [%]	*B*	*M* [%]	*M*	pattern
1364.14	86	195	0	1	E_strain = (24.46:∞]
75.95	67	152	1	14	E_str = (14.90:∞]
29.22	31	71	1	17	E = (147.95:∞]

[a] The growth of the most discriminatory patterns is reported. *B* stands for bioactive/binding and *M* for modeled conformations. For descriptor definitions, see Supporting Information Table 4.

computed patterns by comparing these sets. The three patterns with largest growth rate are reported in Table 2. Only the first two patterns met our predefined support threshold levels (see Methods). Both patterns discriminate effectively, given the magnitude of their growth rates, and clearly reflect the general introduction of ligand strain energy upon complex formation, regardless of the nature of the interactions. The strain energy pattern was dominant with a growth rate of 1364, followed by the bond stretch energy pattern having a growth rate of 76. The third pattern in Table 2 with lower growth indicates that the total potential energy of binding conformations was generally higher than of minimized conformations, which is of course also expected. At the chosen level of support stringency, only patterns were identified that referred to strain or total energy as generally discriminating features. Thus, the global comparison of binding and modeled conformations on the basis of emerging chemical patterns produced meaningful results.

Analysis of Individual Compound Classes. We then attempted to generate discriminatory patterns for each of our

Table 3. Discriminatory Patterns for Three Representative Compound Data Sets

growth	*B* [%]	*B*	*M* [%]	*M*	pattern
					(a) adenosine deaminase
∞	93	14	0	0	E_strain = (24.46:∞]
∞	80	12	0	0	E_str = (14.90:∞]
∞	53	8	0	0	E_tor = (1.88:∞], std_dim2 = (1.81:∞]
∞	53	8	0	0	E_tor = (1.88:∞], pmiY = (1141.06:∞]
∞	53	8	0	0	E_oop = (1.64:∞]
∞	53	8	0	0	E_ang = (17.54:∞]
∞	100	15	0	0	E_strain = (15.69:∞]
∞	87	13	0	0	E_str = (10.26:∞]
∞	73	11	0	0	E_ang = (13.51:∞]
∞	53	8	0	0	E_stb = (0.96:∞]
138.00	100	15	1	1	E = (55.17:∞]
32.20	93	14	3	4	E_oop = (0.25:∞]
27.60	60	9	2	3	E_tor = (1.88:∞], E_vdw = (28.30:72.89]
24.53	53	8	2	3	E_tor = (1.88:∞], PM3_HF = (−85.25:69.90]
24.53	53	8	2	3	E_tor = (1.88:∞], MNDO_HF = (−79.25:169.85]
					(b) ribonuclease A
∞	78	7	0	0	E_vdw = (−∞:28.30]
∞	56	5	0	0	AM1_dipole = (17.34:∞], PM3_Eele = (−755075.90:−335581.90]
∞	56	5	0	0	AM1_dipole = (17.34:∞], MNDO_Eele = (−788514.20:−336721.10]
∞	56	5	0	0	AM1_Eele = (−783413.80:−340732.90], AM1_dipole = (17.34:∞]
∞	78	7	0	0	E_vdw = (−∞:32.76]
∞	67	6	0	0	E_strain = (32.31:∞]
∞	67	6	0	0	E_stb = (−0.36:∞]
∞	56	5	0	0	E_str = (9.28:∞]
∞	56	5	0	0	E = (92.14:∞]
29.33	67	6	2	1	E_strain = (24.46:∞]
6.11	56	5	9	4	AM1_dipole = (17.34:∞], ASA+ = (−∞:227.75], PM3_E = (−124393.40:−95187.02], dipoleZ = (−1.24:∞]
6.11	56	5	9	4	AM1_dipole = (17.34:∞], ASA+ = (−∞:227.75], MNDO_E = (−158562.00:−105111.10], dipoleZ = (−1.24:∞]
4.89	56	5	11	5	dipoleX = (−∞:−1.59], glob = (0.08:0.10]
					(c) trypsin
∞	100	30	0	0	E_strain = (5.11:∞]
∞	63	19	0	0	E_str = (10.03:∞]
∞	63	19	0	0	E_stb = (0.53:∞]
24.25	50	15	2	4	pmiZ = (−∞:16.69]

compound sets. Table 1 reports the number of class-specific patterns that were obtained, given the applied support stringency threshold, which ranged from 5 for trypsin to 197 for thermolysin inhibitors. In most cases, between ∼10 and 30 patterns were obtained. In general, the more diverse the experimental binding conformations are, the fewer widely applicable signature patterns are identified. For three representative sets, the results are reported in Table 3, and for the remaining 15 sets, they are reported in Supporting Information Table 4. As expected, strain energy and related patterns were found in all cases. However, for each of our 18 inhibitor sets, different types of signature patterns also emerged. Most patterns were relatively small including only 1−5 descriptor value ranges, with the exception of the thermolysin set that produced patterns with up to 9 descriptors. This set also generated the largest number of patterns, which is not surprising given the fact that increasing numbers of descriptors lead to an exponential growth in the possible number of patterns. As can be seen in Table 3 and Supporting Information Table 4, the majority of signature patterns had *infinite growth*, which means that they exclusively occurred in bioactive, but not reference conformations. Therefore, these patterns were highly specific. In many instances, patterns consisting of a single energy descriptor were found to be highly discriminatory. Among these were torsion and out-of-plane energy descriptors whose value ranges were indicative of in part significant distortions of inhibitors upon binding. However, patterns containing no energy term descriptors were also found in most cases and usually consisted of combinations of at least two descriptors. For example, the pattern {CASA+ = (1956.91:∞], FCASA− = (1.40:3.21], pmi = (10083.21:∞]} discriminated bioactive factor Xa inhibitor conformations from modeled conformations with an infinite growth rate. This pattern combines charge distributions on the solvent-accessible surface area with the principal moment of inertia (as a descriptor of general mass distribution).

For ligands of the FK506 binding protein, the angle-bending energy and hydrophobic solvent accessible surface area displayed value range preferences for bioactive conformations, but the individual descriptors only partly discriminated between bioactive and modeled conformations. However, a pattern consisting of a combination of the preferred value ranges perfectly classified these conformers. Similar patterns combining energy term descriptors with surface area, charge, or shape features were found to be prominent for many classes. For example, for protein tyrosine phosphatase 1b inhibitors, the pattern {ASA_*P* = (210.39: 282.26], E_ele = (−11.07:∞], E_tor = (1.88:∞]} combines two energy descriptors with polar solvent-accessible surface area and has infinite growth rate. Other examples include ribonuclease A inhibitors where a combination of the dipole

moment and electrostatic energy leads to patterns with infinite growth rate. In some cases, even relatively simple geometric descriptors were found to be highly discriminatory. As an example, for protocatechuate-3,4-dioxygenase, simple shape measures, such as the extension along a molecular axis, effectively distinguished between bioactive and modeled conformations. In this case, the inhibitors are small (substituted benzenes). Upon binding to the enzyme, substituents are forced out of the aromatic ring plane, because of a structural constraint caused by a bound cation, which significantly alters the shape of these small inhibitors. Taken together, pattern analysis for individual compound classes revealed that highly discriminatory patterns of variable composition were identified in each case. These target-specific patterns accurately distinguished between bioactive and modeled conformations of inhibitors.

Signature Patterns and Details of Enzyme−Inhibitor Interactions. Having demonstrated that binding conformation sensitive patterns could be systematically identified for a variety of target sets, we also attempted to go beyond statistical analysis of discriminatory patterns and relate them to details of protein−ligand interactions. Specifically, we were interested in studying how signature patterns could be interpreted in structural terms and, moreover, how they might be used to differentiate between experimental and hypothetical enzyme−inhibitor complexes. These hypothetical complexes were obtained by superposing modeled low-energy conformers onto the crystallographic pose of the corresponding inhibitor. Therefore, the ligand pose was always approximately correct. However, key interactions were expected to differ as a consequence of conformational differences. We reasoned that if this would not be the case, highly discriminatory patterns could hardly emerge. In the following, we present an analysis of three representative cases, inhibitors of adenosine deaminase, ribonuclease A, and trypsin, for which small and intuitive discriminatory patterns were identified. Figure 1 shows the bioactive and modeled conformations of inhibitors used to generate exemplary complexes.

Adenosine deaminases are metalloenzymes that catalyze the deamination reaction of adenosine to inosine. The active site contains a zinc cation that is involved in the activation of a water molecule during catalysis.[19] Table 3a reports the signature patterns for adenosine deaminase inhibitors. Typical strain and bond stretch energy descriptors displayed increased values for binding conformations. In addition, increased out-of-plane and angle-bending energies were characteristic of the majority of binding conformations but none of the modeled conformations. Ring distortions were detected in all inhibitors with available crystallographic binding conformations. Figure 2a shows a crystallographic adenosine deaminase−inhibitor complex. The hypoxanthine ring system is twisted to position the carbonyl oxygen above the ring plane, which results in high out-of-plane and angle-bending energies. In this position, the carbonyl oxygen strongly interacts with the zinc cation and thereby inhibits the enzyme. Figure 2b shows the corresponding hypothetical complex. In the modeled inhibitor, the ring system is planar and not distorted, which prevents the interaction between the carbonyl oxygen and the zinc cation. Figure 2c shows the crystallographic complex formed by a chemically different adenosine deaminase inhibitor that contains three rings. All three ring systems occupy hydrophobic pockets and significantly contribute to the binding affinity.[20] The mode of inhibition is completely distinct in this case. It does not involve complexation of the zinc cation, as discussed above, but rather interactions between a guanidino group of the inhibitor and catalytic residues. In its bound conformation, all three rings of the inhibitor are twisted to varying degrees. The largest contribution to the out-of-plane energy results from a deformation of the benzimidazol-2-amino moiety where the amino substituent is moved out of the ring plane. The ring distortions are an apparent consequence of achieving a degree of shape complementarity while correctly positioning the guanidino group for strong salt bridge and hydrogen bonding interactions with aspartic acid and histidine residues. Figure 2d shows the corresponding hypothetical complex. The modeled ligand with planar ring systems can no longer form the strong interactions via the guanidino group and achieves overall lower shape complementarity, although the benzimidazol ring penetrates deeper into the hydrophobic F2 pocket.

Thus, in the case of adenosine deaminase, signature patterns were identified for a compound set containing chemically distinct inhibitors with different modes of action that discriminated between bioactive and modeled conformations with high accuracy. This was possible because the inhibitors were conformationally perturbed in similar ways upon binding, although their structures and interactions were distinct.

For ribonuclease A inhibitors, we find that strain and bond stretch energy were not the most discriminatory patterns. As reported in Table 3b, other descriptors were found to perfectly discriminate between bioactive and modeled conformations. Signature patterns contained the van der Waals energy as a single descriptor, as well as combinations of the dipole moment and various electrostatic energy term descriptors. The comparison of binding conformations with van der Waals energy values matching the signature pattern and the corresponding modeled conformations revealed that the increase in van der Waals energy in the low-energy conformers resulted from the formation of an intramolecular hydrogen bond involving one of the phosphate groups (Figure 1c) that was not observed in the binding conformation. The phosphate groups of ribonuclease A inhibitors occupy positively charged regions in the active site that accommodate the phosphate groups of the RNA substrate.[21] Figure 2e shows the X-ray structure of a ribonuclease A−inhibitor complex. In its binding conformation, the inhibitor positions the phosphate groups to strongly interact with the histidine and lysine residues in the phosphate binding pockets, thereby achieving charge complementarity. In the corresponding hypothetical complex shown in Figure 2f, the phosphate groups are positioned differently, and many of the interactions seen in the crystal structure can no longer be formed. Here signature patterns correctly detected a conformational artifact that led to a more compact inhibitor structure that would not have been capable of forming the electrostatic interactions within the active site.

Trypsin inhibitors also represent an interesting test case for pattern analysis. Diller and Merz found that trypsin inhibitors were particularly difficult to distinguish from minimized conformations because binding conformations also displayed strong intramolecular interactions between ring

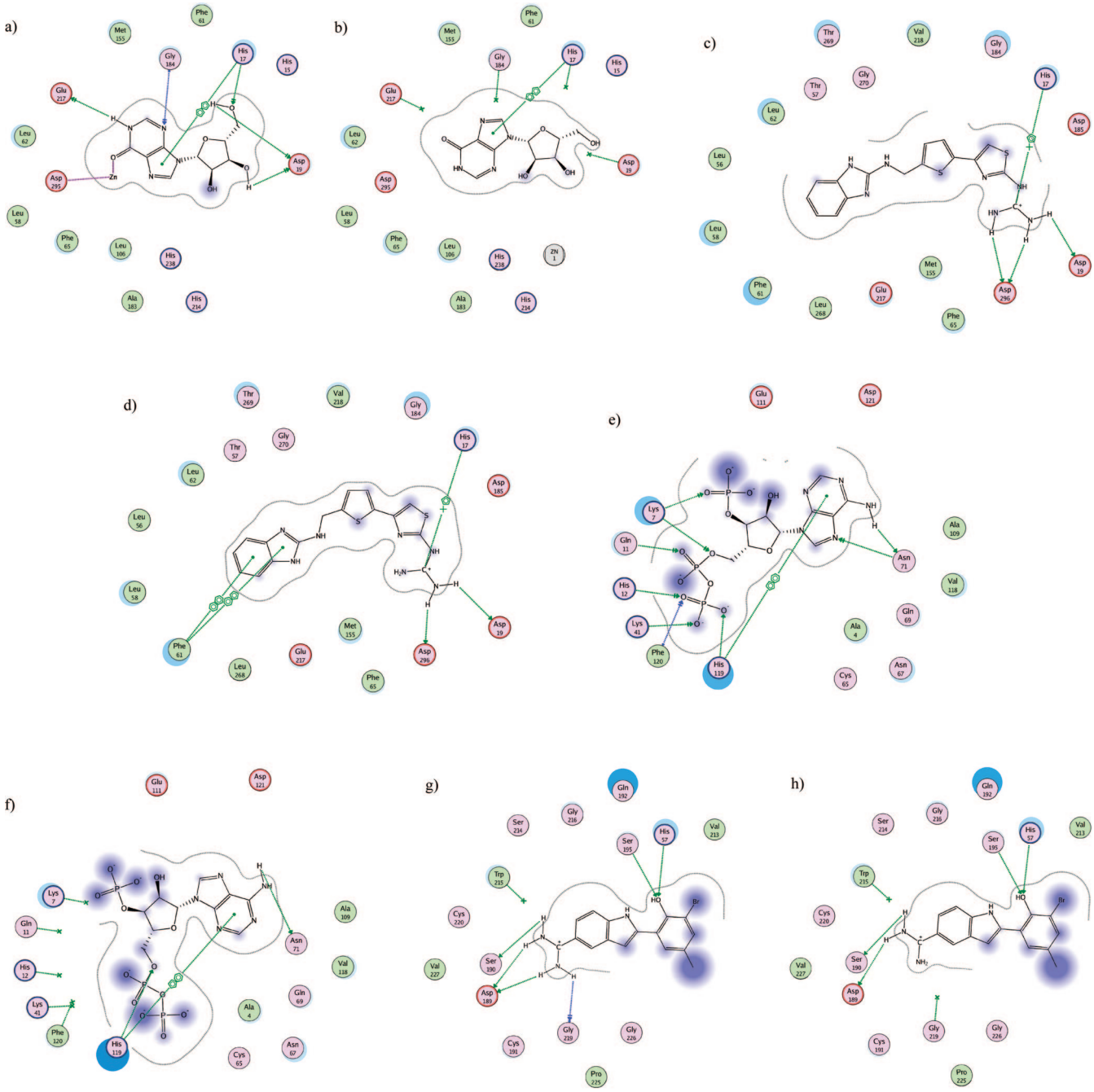

Figure 2. Experimental and modeled enzyme−inhibitor interactions. Interaction diagrams are shown for representative examples of enzyme−inhibitor complexes for which discriminatory patterns were interpreted on the basis of interactions observed in the X-ray structures (see also Figure 1 for ligand conformations). In each case, a modeled conformer was also superposed on the respective inhibitor and consequences for observed enzyme-interactions were studied. In the interaction diagrams, the active site region is delineated as an envelope. Amino acids are color-coded according to polar, negatively charged, and positively charged (pink, red, and blue, respectively) and hydrophpobic (green) character. Solvent-exposed residues and inhibitor atoms have additional shading (light and dark blue, respectively). Hydrogen bonds are drawn as dashed donor−acceptor arrows and are colored green if they involve protein side-chain atoms or blue if they involve backbone atoms. Green dashed lines containing aromatic ring symbols indicate donor interactions with π-electron systems or $\pi-\pi$ interactions. Metal contacts are displayed in magenta. (a) X-ray structure of an adenosine deaminase-inhibitor complex (1fkx). The inhibitor complexes a zinc cation that activates a water molecule during catalysis. The contact is formed through a distorted hypoxanthine ring system. (b) The corresponding hypothetical complex with a low energy conformer. Here the hypoxanthine ring is planar, which prevents the zinc contact, and fewer interactions are formed. (c) Structure of another adenosine deaminase-inhibitor complex (1ndv). Distorted ring systems (with large calculated out-of-plane energy penalty) match hydrophobic binding pockets and present a guanidino group for an array of salt bridge interactions. (d) The corresponding hypothetical complex with a low-energy conformer. In the modeled conformer, the rings are planar, the guanidino group is repositioned, and the salt bridge interaction can no longer be formed. (e) Structure of a ribonuclease A-inhibitor complex (1afk). Binding of the phosphate groups is stabilized by multiple salt bridges and hydrogen bonds. (f) The corresponding hypothetical complex. Many of the key interactions are absent. (g) Trypsin-inhibitor complex (1o3h). The diamino methyl substituent is in a strained equatorial conformation that enables the formation of multiple salt bridge and hydrogen bonding interactions. (h) The corresponding hypothetical complex. Here the diamino methyl group is in energetically preferred axial orientation, which breaks the crystallographic interaction pattern.

systems.[1] However, as shown in Table 3c, in addition to the general increase in strain and bond stretch energy, two discriminatory patterns with single descriptors emerged, accounting for bond stretching/angle-bending cross-term energy and the *z*-component of the principal moment of inertia. Trypsin inhibitors matching these patterns contain a diamino methyl substituent at an indol ring (Figure 1d). In its binding conformation, the diamino methyl substituent is coplanar with the indol ring, which correctly positions the amino groups for well-defined hydrogen bonding and electrostatic interactions (Figure 2g). This conformation has a low *z*-component of the principal moment of inertia because most atoms are positioned in or near the indol ring plane of the molecule. Figure 2h shows the corresponding hypothetical complex. In the modeled conformation, the diamino methyl moiety is rotated such that the amino groups are orthogonal to the indol ring, which increases the value of *z*-component of the principal moment of inertia. However, this orientation would break the network of crystallographic interactions involving the diamino methyl substituent.

These examples illustrate that modeled conformations of the inhibitors studied here could not have been used to correctly predict details of the enzyme−inhibitor interactions, even if pose information was available. However, the analysis shows that key conformational differences can be well rationalized with the aid of discriminatory descriptor patterns.

Conclusions and Outlook. In this study, we have investigated the emerging chemical patterns methodology to systematically detect compound class-specific features that differentiate bioactive and modeled compound conformations. For each of 18 different compound classes, several highly discriminatory patterns were identified based on a pool of 67 3D descriptors. Signature patterns typically consisted of only one to three descriptor value ranges. Consistent with earlier findings, pattern analysis reflected well-known strain energy effects that generally accompany ligand binding. However, signature patterns of individual classes contained more information and accounted for class-specific differences. Signature patterns could be well rationalized on the basis of available structural data and used to distinguish between hypothetical and experimentally observed protein−ligand interactions. In addition to distinguishing between experimental and modeled binding conformations, it is conceivable that descriptor pattern analysis can also be used to identify binding conformations of novel active compounds and targets for which experimental conformations of other ligands are already available. If no experimental conformations are available, binding conformations of other ligands can not be predicted. However, target-specific descriptor patterns can be used as filters to screen conformational ensembles for bioactive conformations. A prerequisite for the success of this analysis is that underlying protein−ligand interactions are similar. In this and our previous study, emerging chemical pattern classifiers of high accuracy were successfully derived from small training sets. This suggests that it should be possible to design pattern filters for identifying bioactive conformations on the basis of relatively few experimentally known ligand conformations, provided the binding modes are overall similar.

Supporting Information Available: Table 1 listing the 3D descriptors used in our study, supplementary methods, including Tables 2 and 3, and Table 4 listing the discriminatory patterns found for all activity classes. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Diller, D. J.; Merz, K. M. Can we separate active from inactive conformations? *J. Comput.-Aided Mol. Des.* **2002**, *16*, 105–112.
(2) Perola, E.; Charifson, P. S. Conformational analysis of druglike molecules bound to proteins: An extensive study of ligand reorganization upon binding. *J. Med. Chem.* **2004**, *47*, 2499–2510.
(3) Agrafiotis, D. K.; Gibbs, A. C.; Zhu, F.; Izrailev, S.; Martin, E. Conformational sampling of bioactive molecules: A comparative study. *J. Chem. Inf. Mod.* **2007**, *47*, 1067–1086.
(4) Stockwell, G. R.; Thornton, J. M. Conformational diversity of ligands bound to proteins. *J. Mol. Biol.* **2006**, *356*, 928–44.
(5) Boström, J.; Norrby, P.-O.; Liljefors, T. Conformational energy penalties of protein-bound ligands. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 383–396.
(6) Nicklaus, M. C.; Wang, S.; Driscoll, J. S.; Milne, G. W. A. Conformational changes of small molecules binding to proteins. *Bioorg. Med. Chem.* **1995**, *3*, 411–428.
(7) Wang, R.; Fang, X.; Lu, Y.; Wang, S. The PDBbind database: Collection of binding affinities for protein-ligand complexes with known three-dimensional structures. *J. Med. Chem.* **2004**, *47*, 2977–2980.
(8) Wang, R.; Fang, X.; Lu, Y.; Yang, C. Y.; Wang, S. The PDBbind database: methodologies and updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.
(9) *Molecular Operating Environment, 2007.09*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
(10) Stewart, J. J. P. MOPAC: A semiempirical molecular orbital program. *J. Comput.-Aided Mol. Des.* **1990**, *4*, 1–103.
(11) Dewar, M. J. S.; Zoebisch, E. G.; Healy, E. F.; Stewart, J. J. P. AM 1: A new general purpose quantum mechanical molecular model. *J. Am. Chem. Soc.* **1985**, *107*, 3902–3909.
(12) Dewar, M. J. S.; Thiel, W. Ground states of molecules. 38. The MNDO method. Approximations and parameters. *J. Am. Chem. Soc.* **1977**, *99*, 4899–4907.
(13) Stewart, J. J. P. Optimization of parameters for semiempirical methods I. Method. *J. Comput. Chem.* **1989**, *10*, 209–220.
(14) Fayyad, U.; Irani, K. On the handling of continuous-valued attributes in decision tree generation. *Machine Learn.* **1992**, *8*, 87–102.
(15) Auer, J.; Bajorath, J. Emerging chemical patterns: a new methodology for molecular classification and compound selection. *J. Chem. Inf. Mod.* **2006**, *46*, 2502–2514.
(16) Dong, G.; Li, J.; Chaudhuri, S.; Madigan, D.; Fayyad, U. Efficient mining of emerging patterns: Discovering trends and differences. In *KDD-99: The Fifth ACM SIGKDD International Conference*; Proceedings of the Fifth ACM SIGKDD International Conference, San Diego, California, 1999; Chaudhuri, S., Madigan, D., Eds.; ACM Press: New York, 1999; pp 43−52.
(17) Auer, J.; Bajorath, J. Simulation of sequential screening experiments using emerging chemical patterns. *Med. Chem.* **2008**, *4*, 80–90.
(18) Clark, A.; Labute, P. 2D depiction of protein-ligand complexes. *J. Chem. Inf. Model.* **2007**, *47*, 1933–1944.
(19) Cristalli, G.; Costanzi, S.; Lambertucci, C.; Lupidi, G.; Vittori, S.; Volpini, R.; Camaioni, E. Adenosine deaminase: Functional implications and different classes of inhibitors. *Med. Res. Rev.* **2001**, *21*, 105–128.
(20) Terasaka, T.; Kinoshita, T.; Kuno, M.; Nakanishi, I. A highly potent non-nucleoside adenosine deaminase inhibitor: Efficient drug discovery by intentional lead hybridization. *J. Am. Chem. Soc.* **2004**, *126*, 34–35.
(21) Yakovlev, G.; Mitkevich, V.; Makarov, A. Ribonuclease inhibitors. *Mol. Biol.* **2006**, *40*, 867–874.

CI8001793