Structural Interaction Fingerprint (SIFt): A Novel Method for Analyzing Three-Dimensional Protein−Ligand Binding Interactions

Zhan Deng, Claudio Chuaqui, and Juswinder Singh*

Department of Structural Informatics, Biogen, Inc., 12 Cambridge Center, Cambridge, Massachusetts 02142

Received July 10, 2003

Representing and understanding the three-dimensional (3D) structural information of protein−ligand complexes is a critical step in the rational drug discovery process. Traditional analysis methods are proving inadequate and inefficient in dealing with the massive amount of structural information being generated from X-ray crystallography, NMR, and in silico approaches such as structure-based docking experiments. Here, we present SIFt (structural interaction fingerprint), a novel method for representing and analyzing 3D protein−ligand binding interactions. Key to this approach is the generation of an interaction fingerprint that translates 3D structural binding information from a protein−ligand complex into a one-dimensional binary string. Each fingerprint represents the "structural interaction profile" of the complex that can be used to organize, analyze, and visualize the rich amount of information encoded in ligand−receptor complexes and also to assist database mining. We have applied SIFt to tackle three common tasks in structure-based drug design. The first involved the analysis and organization of a typical set of results generated from a docking study. Using SIFt, docking poses with similar binding modes were identified, clustered, and subsequently compared with conventional scoring function information. A second application of SIFt was to analyze ∼90 known X-ray crystal structures of protein kinase−inhibitor complexes obtained from the Protein Databank. Using SIFt, we were able to organize the structures and reveal striking similarities and diversity between their small molecule binding interactions. Finally, we have shown how SIFt can be used as an effective molecular filter during the virtual chemical library screening process to select molecules with desirable binding mode(s) and/or desirable interaction patterns with the protein target. In summary, SIFt shows promise to fully leverage the wealth of information being generated in rational drug design.

Introduction

The past decade has witnessed an explosion in the number of three-dimensional (3D) protein−small molecule structures from experimental as well as in silico approaches. At the time this paper was being prepared, the total number of structures deposited into Protein Data Bank (PDB) had exceeded 22 000,[1] with a significant fraction representing protein−small molecule complexes (Battistuz, T. Personal communication). Many more structures are expected to remain undisclosed due to proprietary interests. With the recent development of high-throughput X-ray crystallography, the total number of structures will grow at an even greater speed.[2] In parallel to the growth of experimentally determined structures, a plethora of structural information is also being generated in the rational drug discovery process. A typical virtual chemical library screen could generate a library of structures containing thousands to millions of small molecules docked onto a target protein in silico.[3]

A detailed understanding of intermolecular interactions between proteins and their ligands is of critical importance to structure-based drug design. Traditionally, the interactions between proteins and ligands are rationalized and compared by visually inspecting individual structures on a graphics workstation, sometimes aided by other software tools that generate two-dimensional (2D) schematic representations of the interactions such as LIGPLOT.[4] This traditional approach becomes intractable when the number of complexes to be analyzed is very large, as is the case of the results generated from virtual library screening.

Effective analysis and mining of these virtual structural libraries become a daunting task, as it is impossible to inspect them individually. The large amount of complex structural information requires a new method to help us better analyze the binding interactions between proteins and ligands. Ideally, such a new method should be able to facilitate the following tasks: (i) data visualization, to allow easy interpretation of the binding interactions; (ii) data organization, to organize and cluster the structures in a meaningful way; (iii) data analysis, to enable the comparison and profiling of the binding interactions in different structures; and (iv) data mining, to help search for structures that contain key interactions or specific features. In addition, it is desirable that this method be simple and generic. Here, we present SIFt (structural interaction fingerprint),[5] a simple and robust method for representing and analyzing 3D protein−ligand interactions. Underlying our method is the generation of an interaction fingerprint that converts 3D structural binding information into a one-dimensional (1D) binary string. The representation of the interactions as fingerprints enables clustering,

* To whom correspondence should be addressed. Tel: (617)679-2027. Fax: (617)679-2616. E-mail: juswinder.singh@biogenidec.com.

10.1021/jm030331x CCC: $27.50
Published on Web 12/12/2003

filtering, and profiling of libraries of compounds using approaches that are being widely employed in the field of chemical diversity. We have successfully applied SIFt to analyze large libraries of docking results as well as the crystal structures of the protein kinase family bound to a series of inhibitors.

Materials and Methods

1. Selection of Protein−Ligand Complex Structures. We generated two sets of molecular docking results and used them in our studies. The first set was based upon the crystal structure of p38 in complex with a pyridinyl imidazole inhibitor SB203580 (PDB accession code: 1a9u).[6] The docking program FlexX[7] in Sybyl[8] was used to dock SB203580 onto the crystal structure of p38. In this single ligand study, 100 poses of SB203580 generated by FlexX were retained for subsequent analyses. The ligand binding site was defined using a cutoff radius of 12 Å from the SB203580 ligand (i.e., the conformation in the crystal structure) combined with a core subpocket cutoff distance of 4 Å. The FlexX scoring function was used for scoring the docking. For each ligand being studied, ChemScore,[9] Gscore,[10] PMF Score,[11] Dscore,[12] and Consensus Score[13] were evaluated using the Cscore utility in Sybyl. Consensus scoring attempts to overcome the limitations inherent in any single scoring function by tallying the number of times a ligand was predicted to bind with an enthalpy above a predetermined cutoff threshold across a set of multiple scoring functions. In effect, each function casts a vote as to whether the ligand is a "good" binder, with the Cscore value representing the total number of votes obtained by the ligand. The scoring functions used to derive the consensus score were GScore, PMF Score, DScore, ChemScore, and Fscore[7] as implemented in Sybyl. For each function, a ligand scoring in the upper half of the range of the scores obtained over all ligands was considered to have met the CScore cutoff threshold. Figure 1a shows the 100 poses generated in this experiment. They adopt different orientations and positions in the ATP binding site of the kinase.

The second docking experiment was designed to evaluate the database enrichment potential of SIFt by docking a diverse set of compounds spiked with known actives onto the same target protein structure. To this end, 16 known p38 inhibitors were combined with 1000 small molecules with diverse chemical structures compiled internally. These inhibitors are pyridinylimidazoles and analogues, covering several major p38 inhibitor families reported thus far, as previously discussed by Adams and Lee[14] (also see Supporting Information). These 1016 compounds were docked onto the p38 structure (1a9u) using FlexX distributed across 50 dual processor nodes of a Linux computing farm. For each ligand, 30 different poses generated from the docking experiment were retained, generating a library of 30 480 (30 × 1016) docked ligand structures for subsequent interaction fingerprints analysis. The performance of database enrichment was measured by the enrichment factor (EF),[15] calculated based on the ability of recovering 14 out of 16 (87.5%) known inhibitors. In both docking experiments, 3D conformers of the ligands were generated using OMEGA.[16]

In addition to the virtual structures generated from docking experiments, we also applied SIFt to analyze a family of experimentally determined structures. A panel of 89 X-ray crystal structures of protein kinase−ligand complexes was selected from the PDB.[1] The selection criteria included the following: (i) the structures must contain ligands (either ATP, GTP, or other inhibitors) present in their ATP binding pockets; and (ii) most of the ATP binding site residues are visible and present in the crystal structures. These 89 protein kinase−inhibitor complexes comprise 25 different kinases, covering 14 different protein kinase subfamilies as classified by Hanks and Quinn.[17,18] In all, the kinase structures contain 54 unique compounds representing a variety of chemical structures (see Supporting Information).

2. Construction of Interaction Fingerprints. 2.1. Identification of Ligand Binding Site Residues. The first step in the construction of interaction fingerprints is to identify a list of binding site residues that are common in all complex structures being studied. Here, the ligand binding site is defined as the union of all of the residues that are in contact with any ligand molecules in any of the structures in the group. The resulting panel of ligand binding site residues, which act as a mask covering all of the interactions occurring between the protein and the ligands, is then used as the common reference frame to construct the interactions fingerprints.

The program AREAIMOL[19] of the CCP4 suites[20] was used to identify the protein atoms that are involved in the noncovalent intermolecular interactions with the ligands. AREAIMOL evaluates the solvent accessible area utilizing a probe sphere of 1.4 Å rolling over the van der Waals surface of the protein and the protein−ligand complex. For simplicity, solvent molecules in the PDB files were excluded in our study, although in theory well-ordered solvent molecules can also be included and treated in the same way as protein residues. The ligand binding atoms were identified as nonhydrogen protein atoms exhibiting any solvent accessibility loss upon ligand binding, with the constraint that they are within 4.5 Å from any of the nonhydrogen atoms of that ligand.

In addition, we identified the protein atoms that were involved in hydrogen-bonding interactions with the ligands, using the program HBPLUS[21] with default settings. The program calculated and listed all possible hydrogen bond donor and acceptor pairs in the structure that satisfy predefined geometric criteria. The hydrogen-bonding pairs between protein and ligand were extracted for subsequent analysis.

For a group of structures involving the same target protein (e.g., docking results), the ligand binding site is defined as the list of residues comprising the union of all residues involved in ligand binding over the entire library of structures. For the protein kinase−ligand complex structures, however, as the target proteins involved are different, additional structural and sequence prealignment steps were required.

The crystal structure of murine PKA in complex with ATP and a peptidic inhibitor PKI (PDB accession number: 1ATP)[22] was used as the reference model for sequence and structural alignment. Initial amino acid sequence alignment of the catalytic cores of these kinases was taken from the Protein Kinase Resource (PKR).[18,23] We only focused on a smaller region of the catalytic cores spanning from the Gly rich loop to the catalytic region,[30,31] as we observed that in all structures, almost all of the binding interactions between the protein kinases and the ligands occurred within this region. We then superimposed each protein kinase crystal structure onto 1ATP, using the Homology module in Insight II.[24] The structural alignment was focused primarily on the immediate vicinity of the ATP binding sites. On the basis of the structural alignment results, sequence alignments were carefully checked to make sure that all structurally equivalent residues matched each other in the sequence alignment. The final multiple sequence alignment result was essentially the same as the initial Hanks and Quinn alignment taken from the PKR.[18]

After the alignments, the residues of the nonmurine PKA protein kinases were renumbered and tallied to the murine PKA residue numbering system, resulting in a uniform residue numbering system for all kinases analyzed. Identification of the list of ligand binding sites was carried out as previously described using the new PKA equivalent residue numbers.

2.2. Extraction and Classification of Binding Interactions. After all of the ligand binding site residues were identified and all of the protein−ligand intermolecular interactions were calculated, the next step was to classify these interactions. Seven different types of interactions occurring at each binding residue were extracted and classified from the AREAIMOL and HBPLUS results. They include the following: (i) whether it is in contact with the ligand; (ii) whether any main chain atom is involved in the contact; (iii) whether any side chain atom is involved in the binding; (iv) whether a polar interaction is involved; (v) whether a nonpolar interaction is involved; (vi) whether the residue provides hydrogen bond acceptor(s); and (vii) whether it provides hydrogen bond

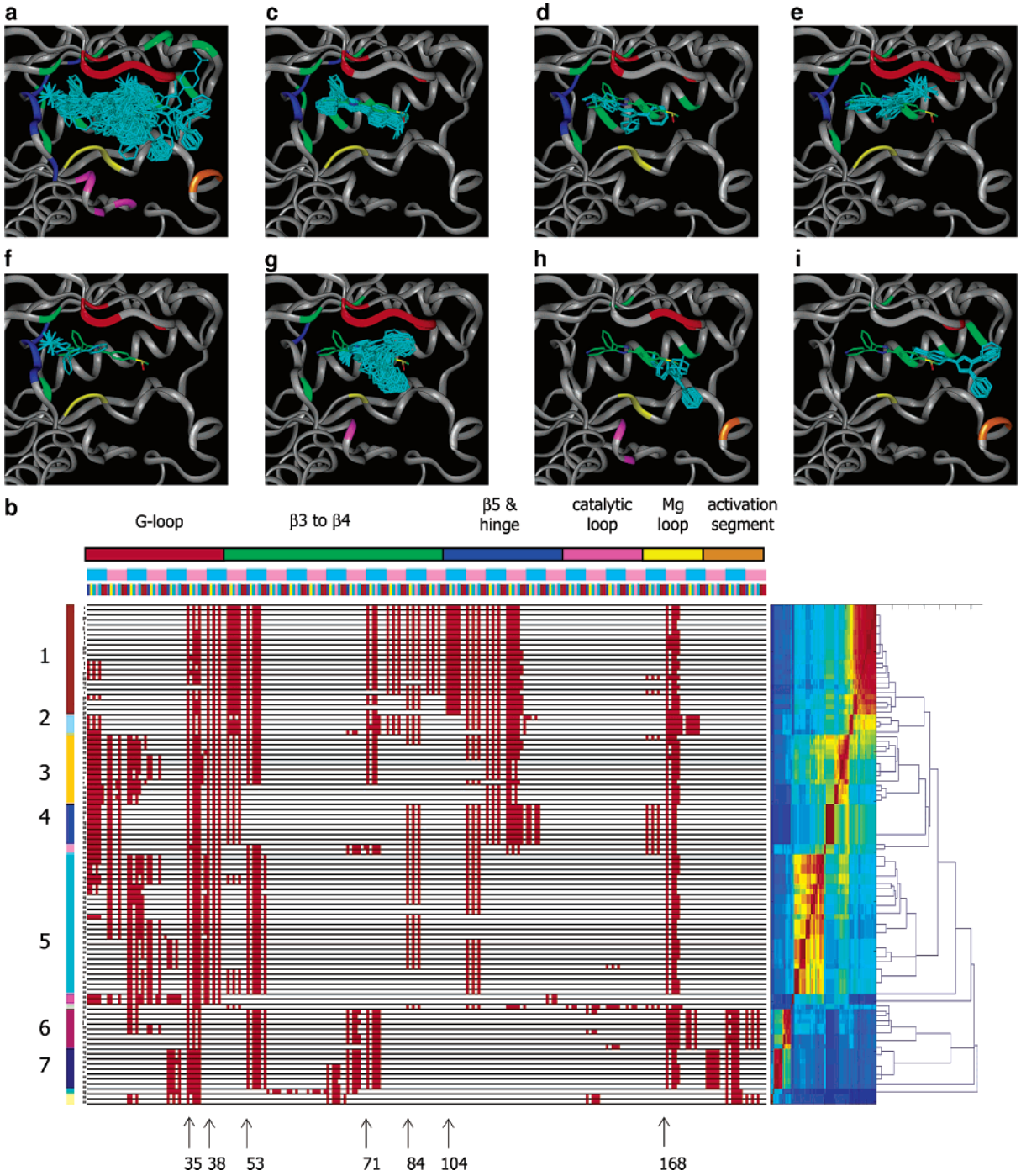

Figure 1. (a) Overlay of 100 different docking poses of SB203580 (shown in cyan stick models) in the vicinity of the target protein human p38 (PDB accession code: 1a9u). p38 is shown as a ribbon model, and the colors represent different subregions of the 34 ligand binding site residues: red, Gly rich loop; green, segment from β3 to β4 (including αC); blue, β5 and hinge region; purple, catalytic loop; yellow, Mg loop; and orange, activation segment. The definitions of the subregions in protein kinase structures are described previously.[30,31] (b) Hierarchical clustering of the SIFts of 100 SB203580 docking poses. Each SIFt is represented as one line in the heat map in the middle of the figure, and only ON-bits (1) are shown as red blocks. The right side of the heat map shows the hierarchical clustering results on the fingerprints, including the dendrogram and the reorganized distance matrix. Colors in the distance matrix correspond to the actual pairwise distance between two SIFts, with dark red being the most similar and dark blue being the least similar. SIFts in the heat map are rearranged according to the order given by hierarchical clustering. The seven major clusters (labeled 1−7) identified from the dendrogram are marked on the left side of the SIFt heat map. The three lines of blocks above the heat map indicate the locations of the corresponding binding site residues and the bits. In the middle line (alternating in blue and pink), each block represents a particular binding site residue, arranged in ascending residue numbers. Within each residue, there are seven different binding bits, represented by seven smaller blocks with different colors in the third line. Also, the residues are grouped into six different regions as described in panel a, as indicated in the first line. Several key residues that make conserved contact interactions with the inhibitors are labeled. (c−i) Overlay of the poses within each of the seven clusters (labeled 1−7), in the same reference frame as panel a. The crystal structure of SB203580 in the 1a9u structure is shown in each figure colored by atom type. Among the 34 binding site residues, only those in contact with the ligands within the respective cluster are colored, using the same color scheme as in panel a.

donor(s). By doing so, each residue was represented by a seven bit long bit string. The whole interaction fingerprint of the complex was finally constructed by sequentially concatenating the binding bit string of each binding site residue together, according to ascendant residue number order. Therefore, interaction fingerprints are of the same length and each bit in the fingerprint represents the presence or absence of a particular interaction at a particular binding site.

3. Analysis of SIFts. 3.1. Measurements of Similarity of Interaction Fingerprints. We have used the Tanimoto coefficient (Tc)[25] as the quantitative measure of bit string similarity. The Tc between two bit strings A and B is defined as:

$$Tc(A,B) = |A \cap B|/|A \cup B|$$

where $|A \cap B|$ is the number of ON bits common in both A and B and $|A \cup B|$ is the number of ON bits present in either A or B. Tanimoto coefficients between random bit strings with a length of 400 bits adopt a near-Gaussian distribution centered at approximately 0.33, with a sigma of about 0.03 (Deng, Z. Unpublished data).

3.2. Hierarchical Clustering of Interaction Fingerprints. Because the interaction fingerprint represents the binding mode of a ligand to a target protein, similar fingerprints imply that the corresponding ligands make similar interactions with the protein. We applied a hierarchical clustering methodology to analyze the fingerprints for each test case. Interaction fingerprints were clustered by using an agglomerative hierarchical clustering approach,[26] applying the Tanimoto coefficients as similarity measurements. Clusters of protein−ligand complex structures were manually selected based on the dendrogram of their interaction fingerprints. Hierarchical clustering analyses were carried out with MATLAB,[27] and all bit string-related computation steps were implemented with the Bit::Vector Perl module[28] in order to achieve fast performance. For more detailed information about the classification of protein kinase structures as well as the 16 known p38 inhibitors used in the database enrichment experiment, see the Supporting Information section.

Results

1. SIFt-Based Analysis of Docking Results. We applied SIFt to analyze the result of a typical docking study. This was comprised of 100 docking poses of a small molecule inhibitor (SB203580) of p38, for which the crystal structure was known (PDB entry 1a9u).[6] The poses adopted diverse binding modes, varied in their orientations and positions relative to the target protein, and were complex to interpret visually (Figure 1a).

A total of 34 protein residues in the vicinity of the ATP binding pocket were identified as the ligand binding site (Figure 1a). These binding site residues are located in different subregions of the kinase structure. Interaction fingerprints were generated for all complexes, each of which was composed of 238 (7×34) binary bits. The hierarchical clustering result of these fingerprints is shown in Figure 1b with the fingerprint Tanimoto similarity matrix represented as a heat map. The dendrogram revealed seven major clusters, labeled 1−7, respectively. Figure 1b shows that the clustering by their SIFt patterns has separated the poses into different groups with distinct binding interactions. Figure 1c−i depicts the structures of each major cluster, put in the same reference frame. Interestingly, each of these seven clusters is comprised of poses having similar binding modes with the receptor; cluster 1 contains molecules similar to the known X-ray crystal structure. Clusters 2−5 are similar in position but represent distinct binding modes that result in dissimilar interactions with the Gly rich loop and the catalytic loop of p38. Finally, clusters 6 and 7 are outside the ATP binding site. Reassuringly, the degree of variation between clusters observed visually in their binding interactions appears to correlate to their distance in the dendrogram. For example, groups 1, 4, 6, and 7 each show very little structural variation, as represented by tight clusters in the dendrogram, whereas groups 3 and 5 show relatively more diversity in their structures as well as in their fingerprints. Furthermore, clusters 1 and 7 have very little in common and are farthest from each other in the dendrogram. In summary, visual inspection confirms that SIFt is useful in separating docking poses into distinct clusters that reveal distinct binding interactions. For more details about the bit values of the SIFt patterns of these seven clusters, see the Supporting Information.

Traditionally, various scoring functions have been used to rank poses from docking studies. Scoring function scores provide an estimate of the binding strength of the compounds in order to identify the potential "good binders" from a large pool of poses, such that a selection of top-scoring compounds derived from a rank ordered list of docked ligands will be enriched with active compounds.[29] We explored how useful scoring functions were at discriminating the poses in the different SIFt clusters (i.e., different binding modes). In Figure 2a, the first SIFt cluster, which is the closest to the true binding conformation, shows a wide range in PMF scores, spanning from the best score (−70) to the worst (−4). In fact, the majority of the poses in this cluster are no better in their PMF scores than those in other SIFt clusters. In addition, the PMF scores for SIFt cluster 2 are just as good as those for cluster 1, even though they adopt different, crystallographically unobserved, interactions with the receptor. Other different clusters also overlap with each other in their docking scores. Clearly, PMF score is a poor scoring function for discriminating compounds with true binding mode and irrelevant poses in our experiment. In an attempt to broaden our analysis of scoring functions, we also examined the consensus scoring function that consists of five commonly used scoring functions (Figure 2b). Many of the poses in clusters 1−3 had high Cscores (3−5), while clusters 3−7 overlap significantly in the range of 0−2. This example further demonstrates the fact that across a range of scoring functions, the energy-based approaches alone were insufficient in distinguishing different binding modes and in isolating those poses corresponding to the observed binding mode.

2. SIFt-Based Analysis of Kinase−Ligand Complex Crystal Structures. We extended the application of the SIFt method to other ensembles of structures involving different proteins and a diverse set of small molecules. We chose 89 known crystal structures of the protein kinase family that had been deposited in the Protein Databank. They represent 14 different protein kinase subfamilies[17,18] and 54 unique kinase small molecule ligands/inhibitors (see Supporting Information). The structure and sequence homology among protein kinases enabled us to analyze these structures using the SIFt-based approach.

A total of 56 residues were identified as the ligand binding site. These residues include (in PKA number-

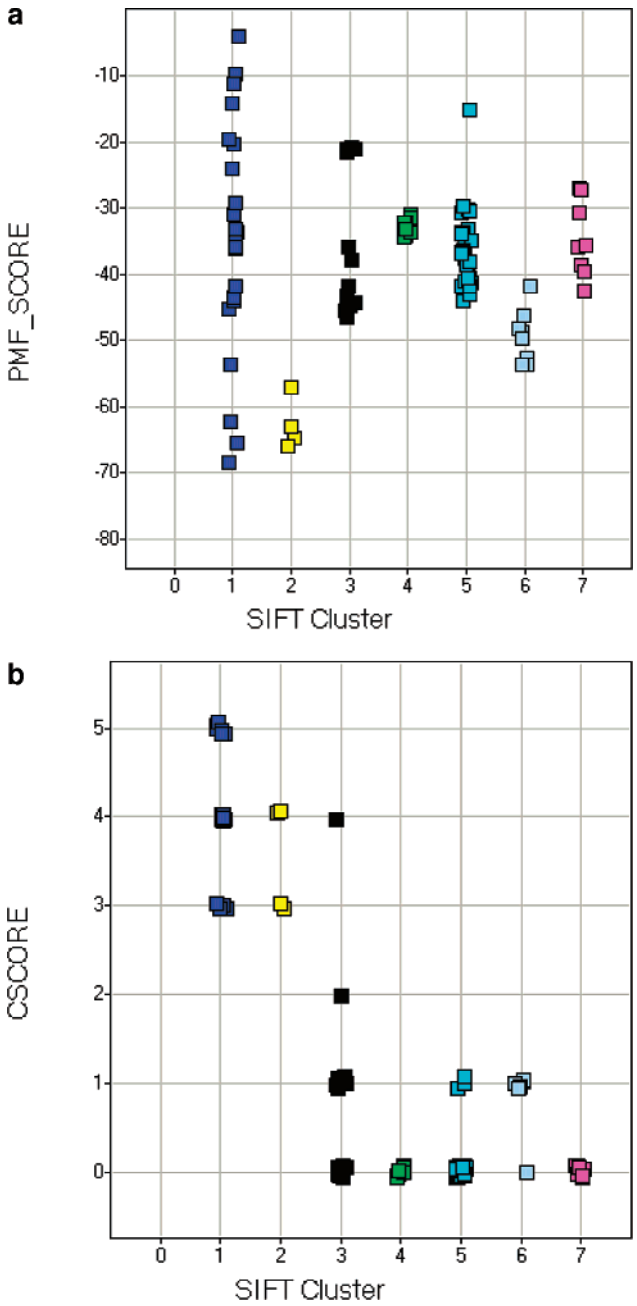

Figure 2. (a) PMF scores as a function of SIFt cluster number. (b) The Consensus score as a function of SIFt cluster number.

ing): 47, 48, 49, 50, 51, 52, 53, 54, 55, 56, 57, 70, 71, 72, 74, 83, 84, 87, 90, 91, 94, 95, 98, 103, 104, 106, 118, 119, 120, 121, 122, 123, 124, 125, 126, 127, 128, 129, 130, 132, 157, 162, 163, 164, 165, 166, 168, 170, 171, 173, 182, 183, 184, 185, 186, and 187. The heat map and the results from hierarchical clustering are shown in Figure 3a. These interaction fingerprints are diverse, reflecting a high degree of variability in their binding interactions. Nevertheless, from the dendrogram, three major clusters can be identified (Figure 3a; also see Supporting Information). Although the results indicate that within each cluster there exists considerable variation in their interaction patterns, these three groups represent three distinct binding modes, as confirmed by careful inspections of their structures (Figure 3b). The first cluster has four members, containing structures of human p38 in complex with four different pyridinyl imidazole inhibitors: SB203580, SB216995, SB220025, and SB218655.[6] The second cluster has 16 members, mostly human CDK2 in complex with different compounds with diverse chemical properties. The third cluster, which does not have a clear-cut boundary, is comprised of approximately 36 structures, and almost all of them are structures of different kinases in complex with ATP or ATP analogue inhibitors (GTP, AMPPNP, AMPPCP, AMP, ADP, etc.). Besides these three major clusters, about one-third of the 89 structures are either singletons or form tiny clusters. Interestingly, the three major clusters represent different grouping examples of protein−ligand complexes—the first one is made up of the same protein and chemically similar compounds; the second group contains the same protein but with a variety of ligands; the third cluster contains different proteins in complex with chemically similar ligands.

Comparison of these fingerprints also revealed interactions that are conserved or highly variable among the structures. For instance, contact interactions with residue 57 (in PKA numbering, within the Gly rich loop) and residue 70 (also in PKA numbering) are strictly conserved among all of the 89 protein kinase−ligand structures. Other highly conserved interactions include contacts with residues 49, 72, 120, 121, 123, 173, 184, etc. (Figure 3a). In contrast, many other interactions are not conserved or only conserved within a particular group. Detailed and systematic comparison of these structural profiles of the ATP binding sites of protein kinases will be presented elsewhere (Deng, Z.; et al. Manuscript in preparation).

3. Data Mining Using SIFt. Our SIFt-based method provides a new and powerful tool for lead discovery and lead optimization, enabling the search for molecules in a chemical database on the basis of expected interaction patterns to a target molecule. To test this application, we performed a virtual screen for a set of 16 known p38 inhibitors spiked into a diverse library of 1000 commercially available compounds. These p38 inhibitors were all ATP competitive inhibitors, and despite representing varied chemical templates, had similarities to the pyridinylimidazole series (i.e., SB203580-like)[14] for which the crystal structure of the complex was known (1a9u).

These inhibitors and the random collection of chemical compounds were docked using FlexX onto the crystal structure of p38 (1a9u), and we assessed how well these known inhibitors could be enriched using commonly used scoring functions. These were then compared with the results from a SIFt-based enrichment involving filtering of the compounds based on their similarities in interaction patterns (measured by Tanimoto coefficient) to SB203580, a known pyridinylimidazole inhibitor of p38 for which the X-ray crystal structure was known. The rationale for SIFt-based enrichment is that these 16 known inhibitors, being diverse analogues of the pyridinylimidazole series, are expected to bind to p38 with similar overall binding modes.

Figure 4a,b and Table 1 show the comparison of the database enrichment performances of the scoring functions with SIFt. ChemScore gave a modest EF of 5.4, and we had to harvest 166 compounds in order to identify 14 of the 16 known p38 inhibitors. PMF was slightly worse than ChemScore, with an EF of 2.0. In

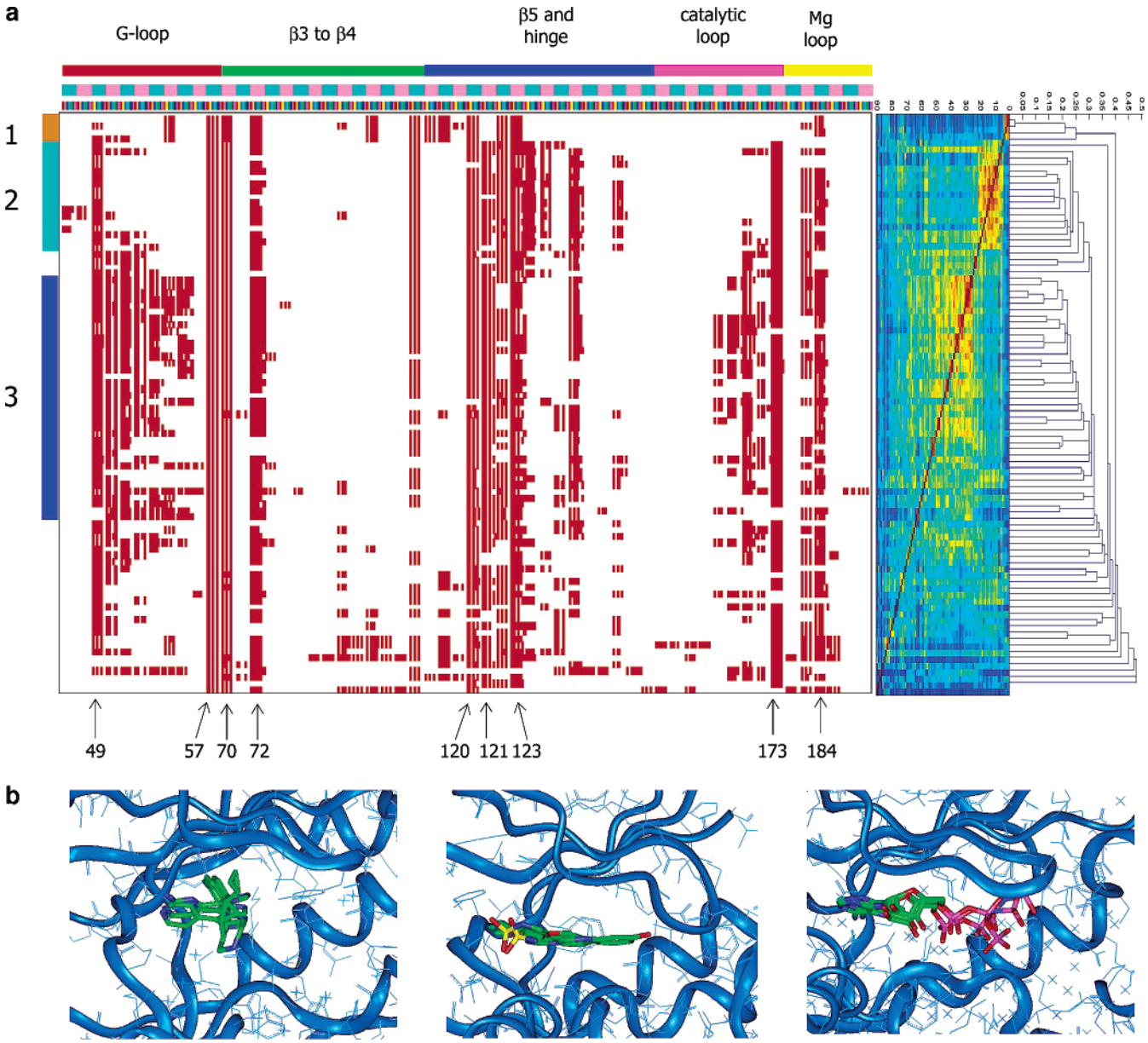

Figure 3. (a) Hierarchical clustering of SIFts of 89 protein kinase crystal structures. On the right are the dendrogram and the corresponding distance matrix map. SIFts are reorganized according to the order given by the dendrogram. Six different regions are labeled above the SIFt heat map. Three major clusters (1−3) are labeled on the left side of the heat map. (b) Comparison of the binding modes of three different kinase clusters (left, cluster 1; middle, cluster 2; right, cluster 3). Three representatives of structures are shown for each cluster.

addition, an analysis of the binding modes of the poses of the enriched p38 inhibitors identified using these scoring functions showed that some of them were highly variable to the known crystal structure of SB203580, despite similarities in functionalities, suggesting that their binding modes obtained by ChemScore or PMF score were incorrect. This implies that the scoring functions were probably performing worse than the EFs were indicating. In contrast, SIFt did pretty well, having to harvest only 24 compounds to be able to identify 14 of the 16 inhibitors, giving an enrichment factor of 37.0. Reassuringly, the highest scoring compound recovered by SIFt was SB203580 upon which the interaction fingerprint used to probe the database was based. Visual inspection of the binding modes of the p38 inhibitors identified using SIFt showed that all of their binding modes were similar to that of SB203580. A combination of SIFt and ChemScore led to a modest increase in enrichment (EF = 42.3).

Discussion and Conclusion

Our results demonstrate that SIFt is a powerful new tool for the visualization, organization, analysis, and data mining of the massive amount of structural information being generated by rational drug design projects. Importantly, it has the potential to significantly reduce the time-consuming user intervention steps currently performed in the analysis of protein−inhibitor complex information.

Virtual chemical library screening is becoming a viable complementary approach to the traditional high-throughput screening for the discovery of novel leads.[32] The virtual screening process generates a huge number of virtual structures, which are very difficult to organize

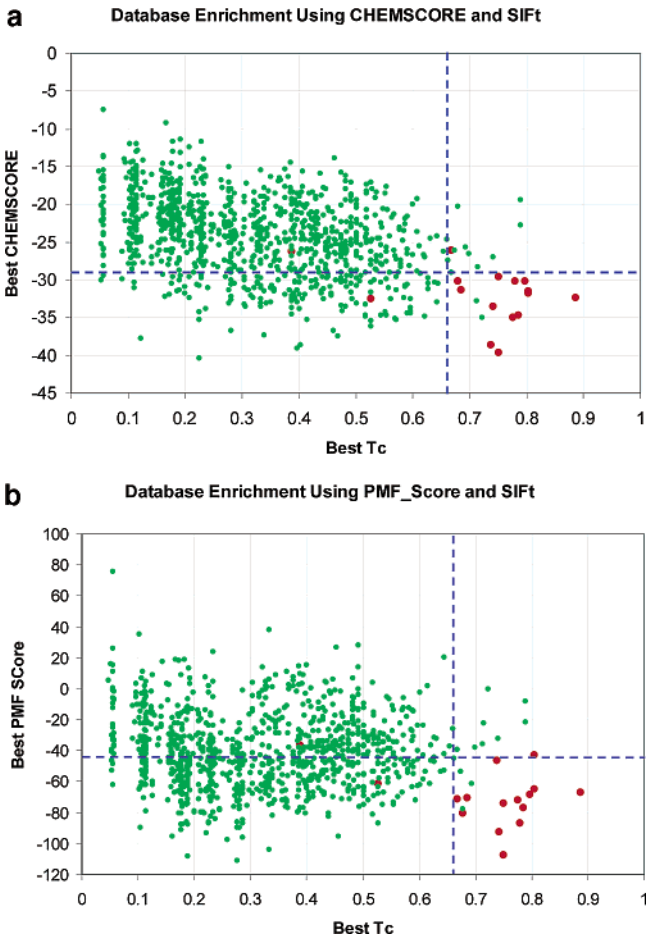

Figure 4. Comparison of database enrichment using SIFt with ChemScore (a) and PMF score (b). Sixteen known p38 inhibitors were diluted in 1000 diverse compounds. The best Tanimoto coefficient (as compared to the crystal structure 1a9u) among 30 docking poses of a compound is plotted against the best ChemScore or PMF score of the same molecule. The red dots represent the 16 known inhibitors, and the green dots represent the 1000 random compounds. The blue-dotted lines indicate the corresponding cutoff scores used to filter the docking poses in order to recover 14 out of 16 known inhibitor (87.5% recovery rate).

Table 1. Comparison of the Database Enrichment Performances of SIFt with ChemScore and PMF Score

filtering method	EF[a]
PMF Score	2.0
ChemScore	5.4
SIFt	37.0
SIFt + ChemScore	42.3

[a] EF is defined as: $EF = (hits_{sampled}/N_{sampled})/(hits_{total}/N_{total})$, where $hits_{sampled}$ is the number of known inhibitors recovered from the sampled fraction of $N_{sampled}$ poses and $hits_{total}$ is the number of known inhibitors present in the whole library of N_{total} compounds.[15] Here, each EF was calculated based on the ability of recovering 14 out of 16 known p38 inhibitors spiked into a random library of 1000 compounds.

with available methods. SIFt can serve as a postdocking molecular organizer and filter, enabling potentially millions of docking poses to be easily reorganized based on their overall binding interaction patterns. Currently, our implementation of the method has allowed us to process and analyze 10 000 docking poses within 4 h using an 800 MHz Pentium III CPU with 1 GB RAM running Linux. This increases 8-fold if we do not compute hydrogen-bonding interactions with HBPLUS. Because the process is highly parallelizable, it could benefit from high-performance computing approaches such as grid computing systems. The linear binary strings generated by SIFt are a simple and highly compact representation of information, which enables us to potentially store and process large numbers of docking results more efficiently using the form of their interaction patterns.

The fingerprint representation of binding interactions makes them amenable to computational approaches that are commonly applied in the analysis of chemical libraries such as clustering analysis for diversity and similarity selection and also key-based searching for database mining. We are currently exploring how to combine chemical diversity-based approaches with the diversity analysis of binding site interactions to explore how both properties relate to each other during a virtual screening experiment (Deng, Z.; Chuaqui, C. Unpublished results).

During lead discovery and lead optimization processes, previously acquired knowledge about how the ligand interacts with the target protein can be used to guide the design and selection of lead compounds for subsequent investigation and refinement. In fact, the potential to quickly compute binding site fingerprints and the low storage requirements makes it possible for a SIFt-based approach to be used to organize and facilitate analysis of the protein databank of X-ray and NMR protein−small molecule complexes. This could serve as a valuable knowledge base of potential desirable interactions within protein families (e.g., protein kinases, aspartyl proteases) that could be used to filter virtual screening results.

Our results highlight the limitations of using scoring function information alone to prioritize the results from a database search. We believe that using both binding interaction constraints and also energy-based constraints are crucial for assigning confidence to the results of a virtual screening experiment. We envision that in the future, the results from SIFt applied to organize X-ray information will be useful in the building of target specific scoring functions that will include both energy-based terms as well as terms that are tailored to the binding site of interest (Deng, Z. Unpublished results).

Our current implementation of SIFt uses seven bits for each binding site residue, representing seven different types of interactions. Although such an implementation has been shown to be able to successfully organize, analyze, and mine a large structural library in a meaningful way, a 7 bit long binary string obviously does not represent all of the binding interactions occurring at a particular residue. The richness of information can be improved by incorporating more bits representing other types of binding interactions, using subresidue portions instead of the whole residue as the basic unit, taking solvent molecules into consideration, or substituting the binary bits with scaled numerical data that reflect the strength and energetics of the interactions. Such an enriched SIFt provides a “higher resolution” picture of the complex. On the other hand, in situations where computational speed is a critical issue, we may construct “lower resolution” SIFts using fewer bits. As a test experiment, we generated SIFts that were merely

comprised of the contact bits, and interestingly, they were able to produce clustering results comparable to that given by 7 bit SIFts (data not shown). Simpler fingerprints give faster performance at the expense of richness of information, and it is especially useful for initial screening of a large structural library, where performance and efficiency are the primary issues. On the other hand, the use of longer fingerprints provides more information at the expense of performance, and it is more useful for detailed structural analysis such as comparing a group of closely related crystal structures. Choosing the proper "resolution" of SIFt is a matter of finding a proper balance between these two competing effects. In addition, we currently treat all seven types of binding bits equally, and consequently, different types of interactions contribute to the final similarity score equally. It is possible to tailor them in a different way by weighting particular types of interaction more heavily than others.

In summary, the SIFt method facilitates and integrates several desirable functionalities including structural data visualization, organization, analysis, and mining, making it an attractive method for analyzing and profiling 3D binding interactions. We envision that SIFt-based methods will be extended to other binary complex systems involving a target molecule and a ligand molecule, including protein−protein interactions, nucleic acid−small molecule interaction, and nucleic acid−protein interactions.

Acknowledgment. We acknowledge the help and assistance of Donovan Chin, Herman Van Vlijmen, Alexey Lugovskoy, and Xin Zhang.

Supporting Information Available: Table of comparison of the SIFt bit strings of seven different poses of SB203580, docked into human p38 structures. Table of 89 crystal structures of protein kinases−ligand complexes and references. Table of 16 known database p38 inhibitors. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) 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.
(2) Blundell, T. L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discovery* **2002**, *1*, 45−54.
(3) Lyne, P. Structure-based virtual screening: an overview. *Drug Discovery Today* **2002**, *7* (20), 1047−1055.
(4) Wallace, A. C.; Laskowski, R. A.; Thornton, J. M. LIGPLOT: a program to generate schematic diagrams of protein−ligand interactions. *Protein Eng.* **1995**, *8*, 127−134.
(5) Deng, Z.; Chuaqui, C.; Singh, J. Patent application pending.
(6) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Yong, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural basis of inhibitor selectivity in MAP kinases. *Structure* **1998**, *6* (9), 1117−1128.
(7) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.
(8) *SYBYL*, version 6.8; Tripos, Inc.: St. Louis, Missouri.
(9) Eldridge, M.; Murray, C. W.; Auton, T. A.; Paolini, G. V.; Lee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(10) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.
(11) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42* (5), 791−804.
(12) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(13) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walter, W. P. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into protein. *J. Med. Chem.* **1999**, *42* (25), 5100−5109.
(14) Adams, J. L.; Lee, D. Recent progress towards the identification of selective inhibitors of serine/threonine protein kinases. *Curr. Opin. Drug Discovery Dev.* **1999**, *2*, 96−109.
(15) Pearlman, D. A.; Charifson, P. S. Improved scoring of ligand-protein interactions using OWFEG free energy grids. *J. Med. Chem.* **2001**, *44*, 502−511.
(16) *OMEGA*; OpenEye Scientific Software, Inc.: Santa Fe, New Mexico.
(17) Hanks, S. K.; Hunter, T. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J.* **1995**, *9*, 576−596.
(18) Hanks, S. K.; Quinn, A. M. Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods Enzymol.* **1991**, *200*, 38−62.
(19) Lee, B.; Richards, F. M. The interpretation of protein structures: estimation of static accessibility. *J. Mol. Biol.* **1971**, *55*, 379−400.
(20) Collaborative Computational Project, Number 4. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **1994**, D*50*, 760−763.
(21) McDonald, I. K.; Thornton, J. M. Satisfying hydrogen bonding potential in proteins. *J. Mol. Biol.* **1994**, *238*, 777−793.
(22) Zheng, J. H.; Tranfny, E. A.; Knighton, D. R.; Xuong, N. H.; Taylor, S. S.; Teneyck, L. F.; Sowadski, J. M. 2.2 Å refined crystal structure of the catalytic subunit of cAMP-dependent protein kinase complexed with Mn-ATP and a peptide inhibitor. *Acta Crystallogr.* **1993**, *D49*, 362−365.
(23) Smith, C. M.; Shindyalov, I. N.; Veretnik, S.; Gribskov, M.; Taylor, S. S.; Ten Eyck, L. F.; Bourne, P. E. The protein kinase resource. *TIBS* **1997**, *22* (11), 444−446.
(24) *Insight II*; Accelrys Inc.: San Diego, CA.
(25) Willet, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(26) Dubes, R.; Jain, A. K. Clustering methodologies in exploratory data analysis. *Adv. Comput.* **1980**, *19*, 113−228.
(27) *MATLAB*, version 6.5; The MathWorks, Inc.: Natick, MA.
(28) Beyer, S. Bit::Vector (Perl module); http://www.engelschall.com/u/sb/download/.
(29) Bissantz, C.; Folkers, G.; Rognan, D. Protein-based virtual screening of chemical databases: 1. Evaluation of different docking/scoring combinations. *J. Med. Chem.* **2000**, *43*, 4759−4767.
(30) Taylor, S. S.; Radzio-Andzelm, E. Three protein kinase structures define a common motif. *Structure* **1994**, *2*, 345−355.
(31) Johnson, D. A.; Akamine, P.; Radzio-Andzelm, E.; Madhusudan; Taylor, S. S. Dynamics of cAMP-dependent protein kinases. *Chem. Rev.* **2001**, *101* (8), 2243−2270.
(32) Singh, J.; Chuaqui, C.; Boriack-Sjodin, P. A.; Lee, W. C.; Ponz, T.; Corbley, M. J.; Cheung, H. K.; Arduini, R. M.; Mead, J. N.; Newman, M. N.; Papadatos, J. L.; Bowes, S.; Josiah, S.; Ling, L. E. Successful shape-based virtual screening: The discovery of a potent inhibitor of the type-I TGFβ receptor kinase TβRI). *Biorg. Med. Chem. Lett.* **2003**, in press.

JM030331X