Interaction Profiles of Protein Kinase−Inhibitor Complexes and Their Application to Virtual Screening

Claudio Chuaqui, Zhan Deng, and Juswinder Singh*

Computational Drug Design Group, Department of Research Informatics, Biogen Idec, Inc., 14 Cambridge Center, Cambridge, Massachusetts 01242

Received August 19, 2004

A major challenge facing structure-based drug discovery efforts is how to leverage the massive amount of experimental (X-ray and NMR) and virtual structural information generated from drug discovery projects. Many important drug targets have large numbers of protein−inhibitor complexes, necessitating tools to compare and contrast their similarities and differences. This information would be valuable for understanding potency and selectivity of inhibitors and could be used to define target constraints to assist virtual screening. We describe a profile-based approach that enables us to capture the conservation of interactions between a set of protein−ligand receptor complexes. The use of profiles provides a sensitive means to compare multiple inhibitors binding to a drug target. We demonstrate the utility of profile-based analysis of small molecule complexes from the protein-kinase family to identify similarities and differences in binding of ATP, p38, and CDK2 compounds to kinases and how these profiles can be applied to differentiate the selectivity of these inhibitors. Importantly, our virtual screening results demonstrate superior enrichment of kinase inhibitors using profile-based methods relative to traditional scoring functions. Interaction-based analysis should provide a valuable tool for understanding inhibitor binding to other important drug targets.

Introduction

The rapid growth in the number of protein−small molecule complexes from X-ray crystallography and NMR has helped drive the success of structure-based drug design (SBDD) and virtual screening (VS) approaches for lead discovery.[1−5] The application of SBDD to the discovery and optimization of drug candidates has led to a tremendous number of protein−ligand complexes for targets[6] including HIV protease, carbonic anhydrase, thrombin, and neuroamidase.

With such large amounts of data being generated, fully leveraging this information hinges on the ability to organize, analyze, and mine the structural data in order to derive knowledge that may be applied to drug discovery. To this end we have developed the structural interaction fingerprint (SIFt) methodology and in a recent paper demonstrated how it could effectively organize, visualize, and analyze protein−ligand complexes.[7] SIFt was also shown to be an effective molecular filter to improve the results obtained from VS.

In this paper we extend the SIFt methodology to derive an interaction profile-based approach we term profile-SIFt, or p-SIFt. The p-SIFt is derived from a collection of SIFts that measures the conservation of interactions observed in clusters of protein−ligand complexes. The p-SIFt approach is analogous to profile-based techniques that have proven to be very useful in the analysis and database mining of groups of protein sequences[8−10] and structures.[11−13] The sequence profile is constructed from a set of multiply aligned sequences or structures of a probe family and is used to identify distant relationships to a database of target proteins. The profile is essentially a sequence position-specific scoring matrix encoding the probability of finding any of the 20 amino acid residues at that position in the target. In the case of p-SIFt, the SIFts derived from a set of probe structures are used to derive a position-dependent profile encoding the probability that a given interaction at that position is present. The probe set of structures may correspond to entire gene families, e.g., kinases, or to subfamilies of structures representing ligands with a particular activity or selectivity profile.

To address the limitations inherent in traditional scoring functions,[14−20] a variety of "knowledge-based" or "target-biased" approaches have been developed that impose constraints based on ligand or receptor pharmacophores thought to be required for activity.[19,21−27] The tractability of knowledge-based VS for lead discovery[3,28−31] is highlighted by recent papers reporting the identification of potent lead (nM) compounds, using receptor-based docking in the case of Chk1 kinase,[22] and ligand-based VS for the discovery of a 25 nM inhibitor against TGFβ-RI.[32] In both of these examples, the success of the VS strategy was dependent on the application of constraints derived from knowledge of how small molecule inhibitors bind at the ATP site of protein kinases. These constraints typically filter virtual libraries based on the presence of kinase binding motifs, or on the ability to satisfy key interactions with the receptor. However, the ability to apply constraints during VS that predict the selectivity of inhibitors for one kinase over another is a much more challenging problem that has not been widely addressed.

The protein kinase family exemplifies the challenges faced with the large amount of structural data being generated not only on specific drug targets but also at the gene family level.[33−35] For example, there exist over

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

10.1021/jm049312t CCC: $30.25
Published on Web 12/16/2004

100 protein kinase small molecule complexes[36] that have been deposited in the public domain, which comprise examples from 34 different kinase family members. Since the majority of kinase inhibitors bind to a conserved ATP site on the enzyme,[36] the ability to understand the selectivity profile for an inhibitor is critical to avoid downstream toxicity issues as well as enable "target-hopping", where an inhibitor to a given kinase is used to discover a lead inhibitor for a new target.[32]

In this paper we describe the application of p-SIFt to analyze the similarities and differences between ATP, p38, and CDK2 inhibitors binding to the protein kinase family. In addition, we demonstrate the ability of p-SIFt to not only enrich for p38 and CDK2 inhibitors, but also, importantly show how it can enrich selectively. Finally, we will demonstrate how p-SIFt can serve as a framework to generate target-specific knowledge-based filters for VS as well as provide an understanding of the interaction patterns responsible for inhibitor selectivity.

Methods

1. Structure Preparation and Docking Methodology. The work reported in this study is based on experimentally determined X-ray cocrystal structures of protein kinases complexed with ATP, ATP analogues, and small molecule inhibitors. A panel of 93 X-ray crystal structures of protein kinase−ligand complexes was selected from the PDB. The selection criteria included the following: (i) the structures must be complexed with small molecules (either ATP, ATP analogues, or inhibitors) present in their ATP binding pockets; and (ii) most of the ATP binding site residues are visible and present in the crystal structures.

We used the crystal structures of p38 in complex with a pyridinyl imidazole inhibitor SB203580 (PDB code 1a9u) and of CDK2 complexed with 4-[3-hydroxyanilino]-6,7-dimethoxyquinazoline (PDB code 1di8) for our docking studies. In each case the ligand-binding site was defined from the bound ligand using a cutoff of 10 Å. Bound waters were removed from the binding sites, and the receptors were protonated at pH 7.4.

The set of known inhibitors of p38 was chosen to span several major p38 inhibitor chemotypes and selected from those reviewed by Adams and Lee.[37] Inhibitors of CDK2 were 54 actives collected from the literature.[27] These known actives for p38 and CDK2 were combined with 1000 small molecules compiled internally. To ensure diversity, the decoy set was selected on the basis of structural and property diversity using the extended connectivity fingerprints (ECFP), molecular weight, and LogP in Pipeline pilot.[38] A 3D version of the ligand database was generated with the program Omega,[39] with options set to generate flexible ring conformers.

The docking program FlexX [40,41] in Sybyl[42] was used to dock onto the crystal structures of p38 and CDK2. In each study 30 ligand poses generated by FlexX were retained for subsequent analyses. The FlexX scoring function was used for scoring the docking.

2. Construction of SIFts. SIFt is a 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. The SIFt protocol and implementation reported previously were used to generate the interaction fingerprints utilized in this study. A uniform list of 56 residues involved in ligand binding was used for both kinase structure analysis and docking studies.

3. Calculation of the p-SIFts. A structural interaction fingerprint profile (p-SIFt) represents the degree to which interactions are conserved across a set of ligand−receptor complexes. The p-SIFt, *P(r)*, is derived from an array, denoted below as **b**, of SIFt patterns, and its derivation from a set of SIFts is shown in Figure 1a. The array has length N for the total number of protein−ligand complexes and width K of SIFt fingerprints bits. The value of each element of *P(r)* is derived by averaging the elements in each column of the SIFt matrix, yielding a numerical interaction frequency that varies from 0 to 1 for unobserved to always present, respectively. The SIFt array, **b,** and resulting *P(r)* are given by,

$$\mathbf{b} = \begin{pmatrix} b_{1,1} & b_{1,2} & b_{1,3} & \cdots & b_{1,K} \\ b_{2,1} & b_{2,2} & b_{2,3} & \cdots & b_{2,K} \\ & & \vdots & & \\ b_{N,1} & b_{N,2} & b_{N,3} & \cdots & b_{N,K} \end{pmatrix}$$

and

$$P(r) = [P_1 \quad P_2 \quad P_3 \quad P_4 \quad P_K]$$

where $b_{i,r}$ is the binary bit value in the SIFt $i=1,N$ at position $r=1,K$. The value in the p-SIFt at position r is given by,

$$\mathrm{P}(r) = \sum_{i=1}^{N} b_{i,r}/N$$

4. Measurement of Similarity between SIFts and/or p-SIFts. We have used the Tanimoto[43] coefficient to measure the similarity between two SIFts, between two p-SIFts, and between a SIFt and a p-SIFt. The Tanimoto coefficient (T_c) between a p-SIFt, *P(r)*, and SIFt, **a**, is defined as,

$$T_c = \frac{\sum_{i=1}^{K} P_i \cdot a_i}{\sum_{i=1}^{K} P_i^2 + \sum_{i=1}^{K} a_i^2 - \sum_{i=1}^{K} P_i \cdot a_i}$$

where K denotes the number of bits defining the SIFt and p-SIFt. Other similarity metrics were tested, including the cosine coefficient, Pearson's correlation coefficient, Dice coefficient, and city block distance, and were found to not affect the relative rankings.[44−47] As has been described previously, a set of SIFt patterns can be clustered using the Tanimoto similarity measure by applying standard hierarchical clustering algorithms.[45,48]

The statistical Z-score was employed to measure how significant the similarity between a SIFt and a target p-SIFt (i.e., a group of structures) is above a certain

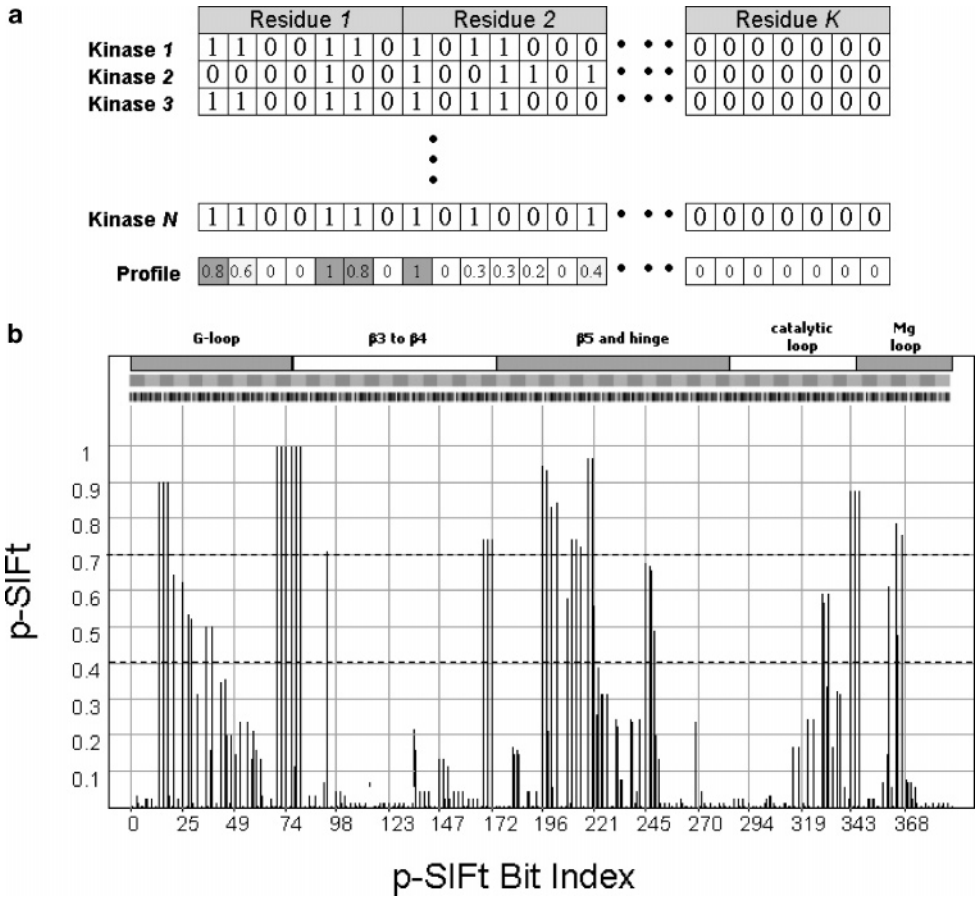

Figure 1. The procedure used to generate a p-SIFt from a set of SIFts is illustrated in panel (a). The profile shown in panel (b) corresponds to the p-SIFt generated from the 93 kinase structures using all 7 bits to compute the SIFts, where the *x*-axis denotes the bit index. The p-SIFt is annotated with a topmost bar delineating the general kinase structural features for that portion of the fingerprint; the bar below consists of alternating blocks corresponding to each residue (site in the uniform PKA numbering scheme) in the kinase used to construct the fingerprint; finally the third bar consists of blocks for each bit representing the interaction features at that site.

background. The *Z*-score is an indication of how many standard deviations an observation differs from the mean. In the present context, the *Z*-score indicates how far and in what direction the Tanimoto coefficient deviates from the background distribution's mean value, expressed in units of this distribution's standard deviation. A *Z*-score of 1.0 and 2.0 corresponds to a level of significance (one-tailed probability) of 0.16 and 0.02, respectively. For our purposes the *Z*-score is defined as,

$$Z_{\text{target}} = \frac{x_{target} - \langle x_b \rangle}{\sigma_b}$$

where target = CDK2 or p38, x_{target} is the Tanimoto coefficient of the SIFt against the target p-SIFt, and $\langle x_b \rangle$ and σ_b are the mean and standard deviation of the Tanimoto coefficients of all the SIFts in the background set, respectively, against the same target p-SIFt.

An artificial kinase set of SIFt patterns was used to define a reference background distribution upon which the comparisons were based. For the kinase crystal structures analysis, we generated a set of dummy SIFts based on the distribution given by an all kinase p-SIFt. The p-SIFt from all 93 kinase crystal structures was calculated first (shown in Figure 1b). We then generated a background set of 2000 bit-strings of the same length, such that each position within these bit-strings was randomly assigned either 1 or 0, with a probability of assigning a value of 1 equal to the value in the all kinase p-SIFt. For the database enrichment experiment, we used all the docking poses as the background set.

5. Virtual Screening Protocol and Postprocessing. In this paper, we explored three strategies for VS postprocessing. For all three strategies a list of docking poses was generated using the program FlexX, and the top 30 poses were retained using the FlexX scoring function.[40] The output obtained from docking *N* ligands is then an *N*x*M* matrix consisting of *M* poses (here, *M* = 30) for each docked ligand. The aim of postprocessing the docking results was to arrive at a rank ordered list of *N* ligands, consisting of a single pose per ligand, which is enriched with actives. The degree of enrichment will depend on the success of the postprocessing strategy.

All three of our strategies required the selection of a single pose per ligand and then subsequent ranking of those ligands. The first approach utilized energy-based scoring functions for both selecting the top pose per ligand, and for ordering the ligand list (traditional scoring). The second approach involved using a p-SIFt (instead of an energy-function) to both select the single best pose per ligand and order the ligand list (p-SIFt scoring). The final approach was a hybrid of the two approaches, in which the p-SIFt was used to filter out undesirable poses, and then an energy-based scoring

Table 1. Postprocessing Schemes Applied in this Paper To Score the Ligand Poses Generated from the Docking Experiments[a]

	Postprocessing Method		
	traditional scoring	p-SIFt scoring	hybrid scoring
A. rescoring	ChemScore, Gscore, PMF Score, Dscore, and Consensus Score	Z_{target}	ChemScore, Gscore, PMF Score, Dscore, and Consensus Score Z_{target}
B. filtering	none	none	$Z_{CDK2} < 4.5$ $Z_{p38} < 5.0$ canonical interactions
C. final pose selection	scoring function	Z_{target}	scoring function
D. ligand ranking	scoring function	Z_{target}	scoring function

[a] In Table 1, scoring function refers to either ChemScore,[55] Gscore,[56] PMF Score,[57] Dscore,[58] or Consensus Score.[59] For both the traditional scoring and hybrid scoring schemes, the same scoring function was used for final pose selection (step C) and ligand ranking (step D).

Table 2. Summary of the Raw Frequencies Observed for Contact Interactions Only, Where Residues Having a Frequency Greater than 0.4 for Any Subgroup Are Listed[a]

	Raw Interaction Frequency						
PKA#	All	ATPg	CDK2	p38	non-ATP	2-Structure	Interaction Context
49	0.9	0.9	0.9	0.4	0.9	Gly-Rich Loop	ATP; Hydrophobic contact with Adenine
50	0.6	0.9	0.3	0.2	0.5	Gly-Rich Loop	ATP; Ribose
51	0.5	0.7	0.3	0.1	0.4	Gly-Rich Loop	ATP; Ribose
52	0.5	0.9	0.4	0.0	0.3	Gly-Rich Loop	ATP; Phosphate
53	0.4	0.7	0.2	0.1	0.1	Gly-Rich Loop	ATP; Phosphate
54	0.3	0.5	0.2	0.9	0.2	Gly-Rich Loop	ATP; Phosphate
55	0.2	0.5	0.1	0.0	0.1	Gly-Rich Loop	ATP; Phosphate
57	1.0	1.0	0.7	0.8	1.0	Gly-Rich Loop	ATP; Hydrophobic contact with Adenine, Ribose, Phosphate
70	1.0	1.0	1.0	1.0	1.0	β3	ATP; Hydrophobic contact with Adenine
72	0.8	0.9	0.7	1.0	0.8	β3	ATP; Phosphate
95	0.1	0.0	0.0	0.6	0.2	αC	Hydrophobic pocket
104	0.7	0.7	0.7	0.8	0.8	Loop- αC-α4	ATP; Hydrophobic contact with Adenine
106	0.1	0.0	0.0	0.4	0.1	Loop- αC-α4	Hydrophobic pocket
118	0.2	0.0	0.0	1.0	0.3	β5	Hydrophobic pocket
119	0.0	0.0	0.0	0.7	0.1	β5	Hydrophobic pocket
120	0.9	0.9	0.9	1.0	1.0	β5	Gatekeeper
121	0.8	0.9	1.0	1.0	0.8	β5	ATP; Hydrogen bond with Adenine
122	0.7	0.6	1.0	1.0	0.8	β5	ATP; Hydrophobic contact Adenine
123	1.0	1.0	1.0	1.0	1.0	hinge	ATP; Hydrogen bond Adenine
124	0.3	0.0	0.6	0.4	0.5	hinge	ATP; Adenine water mediated interaction
125	0.2	0.0	0.5	0.4	0.4	hinge	
127	0.7	0.8	0.9	0.3	0.6	hinge	ATP; Ribose
130	0.3	0.2	0.5	0.0	0.4	hinge	ATP; Ribose water mediated interaction
168	0.2	0.5	0.2	0.0	0.1	Loop- β6-β7	
170	0.6	0.8	0.4	0.2	0.4	Loop- β6-β7	ATP; Ribose
171	0.3	0.4	0.4	0.0	0.3	Loop- β6-β7	
173	0.9	0.9	1.0	0.3	0.9	Loop- β6-β7	
182	0.0	0.0	0.0	0.0	0.0	β8	ATP; contact with Mg-Loop region
183	0.6	0.5	0.3	0.4	0.7	β8	ATP; Hydrophobic contact with Mg-Loop region
184	0.8	0.9	0.8	0.7	0.8	β8	ATP; contact with Mg-Loop region

[a] Residues are colored according to interaction conservation: conserved ≥ 0.7 (green), 0.4 ≤ intermediate < 0.7 (yellow), variable < 0.4 (red). Highlighted cells in the annotation columns indicate that the frequency was defined as conserved (≥0.7) for all subgroups independently. Wherever possible, information on the context of the interaction in binding ATP or inhibitors is included as an annotation.

function was used to select the best pose per ligand and to create an ordered list of ligands (hybrid scoring).

In all three cases, the overall postprocessing scheme used to triage ligand placements, or poses, generated from docking consisted of four general steps, namely,

A. rescoring: each pose generated (*N*x30) is scored using standard scoring functions and p-SIFts

B. filtering: unrealistic poses are removed from the pool to be considered further.

C. final pose selection: a single pose per ligand is selected.

D. ranking of ligand list: the N ligands are rank ordered.

The three scoring and postprocessing strategies we have applied utilized different strategies to carry out steps A−D as summarized in Table 1. For the traditional scoring and hybrid scoring schemes, docked poses were rescored using several widely applied scoring functions computed using the Cscore[49] utility in Sybyl.[42] For the p-SIFt scoring and hybrid scoring protocols we also computed Z between the SIFt for the pose and the target p-SIFt, Z_{target} , where target = CDK2 or p38.

The filtering step B applied in the hybrid scoring scheme involved filtering out any poses having $Z_{CDK2} < 4.5$ and $Z_{p38} < 5.0$, for VS against CDK2 or p38, respectively. The Z_{target} cutoffs were chosen to be at the lowest value of the Z_{target} distribution observed for the CDK2 and p38 X-ray structures. In addition, a canonical interaction filter was applied to each pose such that SIFts not satisfying the subset of interactions having an interaction frequency of 1 in the all-kinase p-SIFt, shown in Figure 1b, were filtered out (see section 2).

It is during the filtering step that incorrect ligand poses can be eliminated from the pool of poses that will

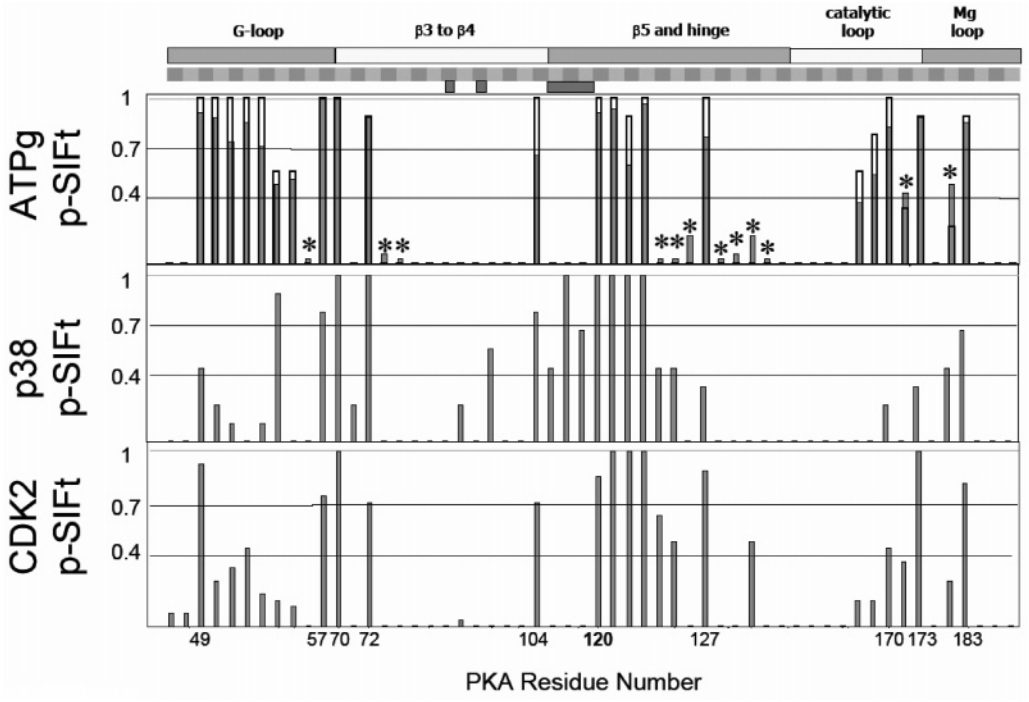

Figure 2. The contact-only p-SIFts for ATPg (top panel), p38 (middle panel), and CDK2 (bottom panel) are plotted as a function of PKA residue numbering. The unshaded outline shown in the ATPg panel corresponds to the p-SIFt derived from the nine ATP-only structures. A decrease in conservation when ATP analogues are introduced is clearly visible by the proportion of the bar that is shaded. Interactions that exploited by ATP analogues beyond those utilized by ATP alone are labeled by (*) for clarity. The blocks above the ATPg p-SIFts denote residues making up the hydrophobic pocket of the kinase.

be considered for final selection. The aim is therefore to reduce the number of false positive poses while retaining the plausible true positive poses. The filtering step is optional and for comparison purposes was omitted in order to generate results based only on functions (traditional scoring) and only on p-SIFTs (p-SIFt scoring).

Step C involves selecting a single pose per ligand from the set of poses that have passed all of the filters, if any, applied in step B. Enrichment curves and factors were computed by rank ordering (step D) the final set of ligand poses using the schemes outlined in Table 1.

Results

1. Implementation of the SIFt Methodology for the Kinase Family. The work presented in this paper is based on 93 X-ray structures representing 21 different protein kinases complexed with ATP, ATP analogues, and a variety of small molecule inhibitors. Of the 56 ligand binding site residues used to construct the SIFts, those playing a significant role in interactions with ligands are listed in Table 2, along with their uniform protein kinase A (PKA) residue numbering.

The results of the hierarchical clustering of SIFts computed for the 93 kinase structures were described in the original SIFt paper[7] and revealed three major clusters representing three dominant interaction patterns present in the ligand−kinase complexes. Cluster 1 is composed of nine structures of small molecule inhibitors interacting with p38 kinase (herein referred to as the p38 cluster). Similarly, Cluster 2 is composed of 20 structures for complexes involving inhibitors of CDK2 kinase (denoted as the CDK2 cluster). The largest distinct group, Cluster 3, is made up of nine ATP and 16 ATP analogues complexed with different kinases, which will be termed the ATP-group (ATPg) cluster. The remaining 42% of the structures do not belong to any particular cluster. It is noteworthy that the hierarchical clustering procedure, based solely on ligand−receptor interaction features, is able to group structures into meaningful clusters where variable ligands have similar interactions with a fixed receptor (p38 and CDK2 clusters) and where very similar ligands interact in a highly conserved way with a diverse set of kinase receptors (ATPg cluster).

2. Interaction Profiles and Profile−Profile Analysis. Each SIFt cluster represents a set of structures having similar conserved and variable interactions. One approach to represent the degree of interaction conservation for any set of structures is to define a p-SIFt from the SIFts of those structures, where the p-SIFt measures the degree of conservation of an interaction feature at that residue. The p-SIFts may be derived using a reduced set of interaction features to represent each interaction. Thus, while the standard SIFts utilizes 7 bits to characterize the interaction at each residue, a simplified p-SIFt may be derived from only the interaction frequencies of the contact bit at each residue.

To simplify the analyses, results presented in this section were based on contact-only p-SIFts. The upper panel of Figure 1 illustrates the general methodology used to derive a p-SIFt from a set of SIFts.

As an initial application, p-SIFts provide a useful tool to overview the interaction patterns observed between ligands and protein kinases. For this purpose, it is convenient to define categories from the contact-only p-SIFts to characterize the observed interactions, e.g., conserved ≥ 0.7, $0.4 \leq$ intermediate < 0.7, variable < 0.4, as denoted by dashed lines on the plot in Figure 1b. It should be noted that the p-SIFts themselves are

not particularly sensitive to minor variations in the cutoffs used for binning the interaction frequencies. The overall distribution of conserved, intermediate, and variable interactions observed overall and for the ATPg, p38, and CDK2 clusters are summarized in Table 2.

Conservation in ATP and ATP Analogues. The 25 members of the ATPg cluster consist of nine structures of ATP complexed with three different kinases and 16 structures of ATP analogues complexed with six kinases. The ATPg p-SIFt computed from the ATPg cluster SIFts is shown in the top panel of Figure 2. For comparison we have also plotted the p-SIFt derived using only the 9 ATP structures in the ATPg cluster. For the nine ATP complexes, 18 out of 23 contacts are classified as conserved between the kinases and the ribose, triphosphate, and adenine moieties. Moreover, for the nine ATP structures there are no completely variable positions. Interestingly, even for these ATP-only structures, four interactions lie in the intermediate conservation range. Interactions between the γ-phosphate and residues 54 and 55, making up the tip of the glycine-rich loop in the kinases, are dependent on the conformation of this flexible region of the binding site and are observed only in approximately half of the structures. Contact between the β-phosphate of ATP and residue 171 is primarily determined by the conformation of the ATP phosphate groups. In approximately 60% of the structures, the $\alpha-\beta$ phosphate pyrophosphate bond is rotated such that the β-phosphate is oriented away from residue 171 and toward the glycine-rich loop (Figure 3; PDB code 1atp). It is noteworthy that in several of these structures a water molecule is observed to take the place of the rotated β-phosphate and form water-mediated interaction between ATP and residue 171.[50] Finally, contact between the adenine ring of ATP and residue 183 is largely a function of the side-chain identity. No contact is observed for the ATP structures that have Ala at this position (50%) whereas Thr and Val side-chains are able to contact the adenine ring either directly or via water-mediated interaction.

When the ATP analogues are considered in addition to the ATP complexes, the degree of variability increases. In particular, interactions with residues 104, 122, and 168 shift from conserved to variable. The extent of variability is clear when the ATPg p-SIFt is compared to the ATP only p-SIFt, as shown in Figure 2. The contacts that are not fully conserved for the ATPg cluster are colored yellow in Figure 3a and fall into the intermediate (~21%) and variable (~33%) ranges. Nevertheless, the ATPg p-SIFt reveals a high degree of interaction conservation as annotated in Table 2 and colored green in Figure 3a. Of the 33 contacts observed across the ATPg, ~46% are classified as conserved (see Table 3). The patterns of conserved interactions for ATPg ligands define an ATP-like binding signature and provide a baseline for comparison when analyzing non-ATP small molecule inhibitors

Non-ATP p-SIFts and Difference Profiles. The contact p-SIFts derived for the ATPg, CDK2, and P38 clusters plotted in Figure 2 measure the degree of interaction conservation for each group of structures. From the p-SIFts, it is evident that CDK2 and p38 inhibitors share some common binding interactions as observed between ATP and some regions of the kinase domain while displaying marked differences in others. A convenient way to compare directly the p-SIFts is to define a difference profile computed by the direct subtraction of one p-SIFt from another. The difference profiles provide insight into how the interaction patterns observed for known kinase inhibitors differ from those detailed above for ATP. To this end, we have defined difference profiles p38-ATPg, p38-CDK2, and CDK2-ATPg, plotted in Figure 4.

For the p38-ATPg and p38-CDK2 difference profiles, the key distinctions are determined in part by the identity of the residue at position 120. Referred to as the "gatekeeper" residue, it controls the relative access to the hydrophobic pocket of the ATP site, a region not occupied by ATP. Bulky residues at position 120, such as the Phe in CDK2, restrict access to the hydrophobic

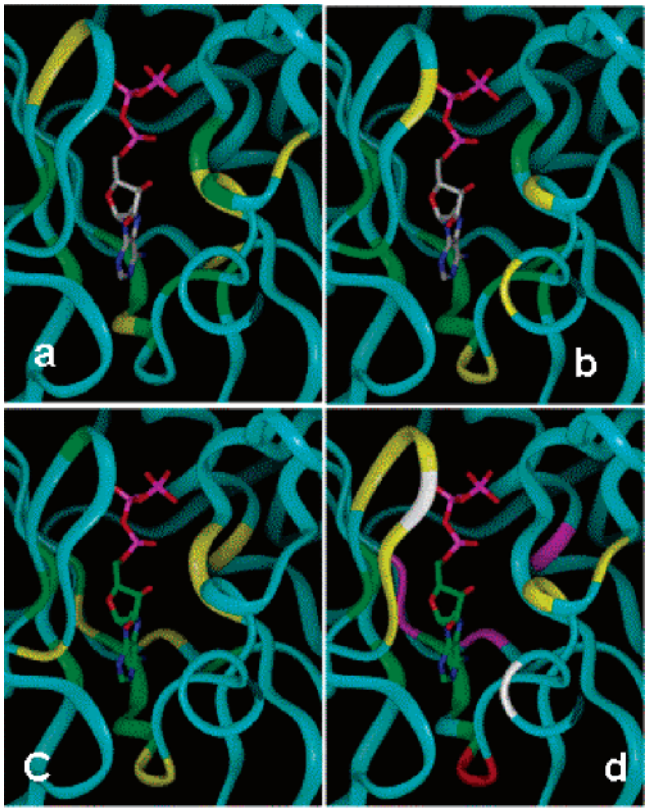

Figure 3. In panels a−c the binned contact-only p-SIFts for ATPg, CDK2, and p38, respectively, are mapped onto the structure of the complex between ATP and PKA (PDB code 1atp) using the values and color scheme introduced in Table 2. Panel d highlights key areas of difference in the interaction patterns observed for ATP, CDK2, and p38 identified from the difference profiles.

Table 3. Summary of Conserved, Intermediate, and Variable Interactions Observed across All of the 93 Kinase Structures and for Each of the ATPg, P38, and CDK2 Structure Clusters[a]

		cluster		
	ALL	ATPg	P38	CDK2
contact residues	56	33	26	29
conserved	11 (19.6%)	15 (45.5%)	10 (38.5%)	12 (41.4%)
intermediate	6 (10.7%)	7 (21.2%)	8 (30.8%)	5 (17.2%)
variable	39 (69.6%)	11 (33.3%)	8 (30.8%)	12 (41.4%)
unique conserved	5 (8.9%)	9 (27.3%)	4 (15.4%)	6 (20.7%)

[a] Total number of residues interacting with ligands in each group is denoted as "contact residues". The number of conserved interaction beyond the canonical set observed for all ligands appears in the row labeled "unique conserved".

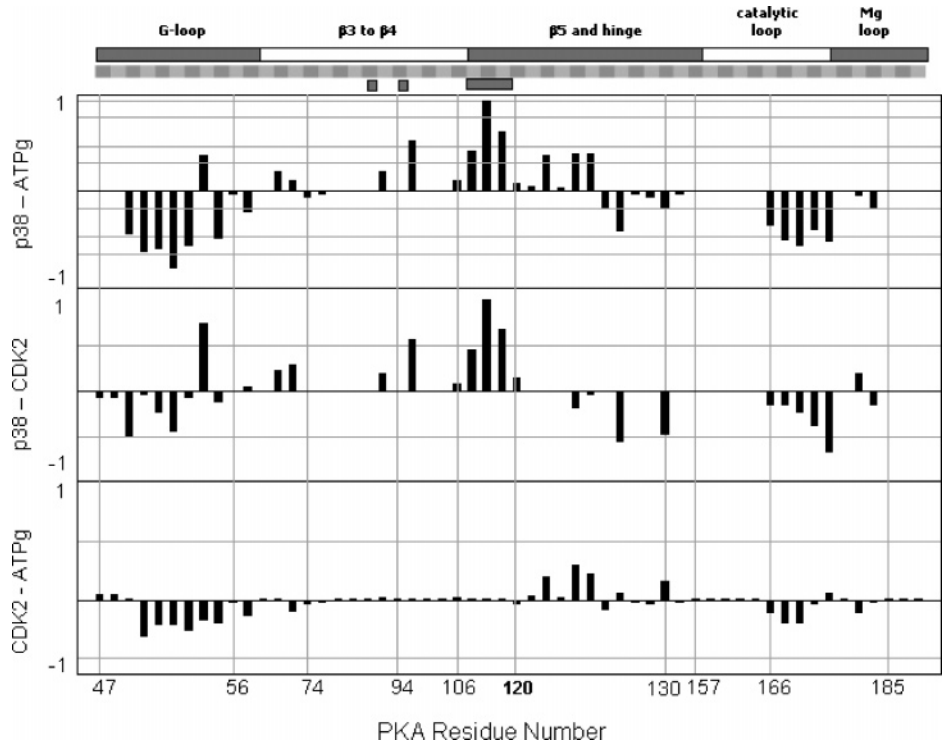

Figure 4. The contact-only difference profiles between p38-ATPg (top panel), p38-CDK2 (middle panel), and CDK-ATPg (bottom panel). The difference plots range from −1 to 1, where a value of 0 indicates that the interaction is conserved to the same degree in the two sets of structures, whereas a value of −1 or 1 denotes that a conserved interaction in one set of structures is not conserved in the other.

pocket, limiting the contacts available to a putative inhibitor. The small Thr "gatekeeper" in p38 renders the residues making up the hydrophobic pocket accessible to small molecule inhibitors. The fact that small molecule inhibitors of p38 exploit these interactions is clearly evident from the p38 p-SIFt (Figure 2), which indicates a set of intermediate and conserved interactions corresponding to hydrophobic pocket residues colored magenta in Figure 3(c). The contrast in interaction with the hydrophobic pocket observed between p38, ATPg, and CDK2 is clearly delineated by the distinct positive differences visible in the p38-ATP and p38-CDK2 difference profiles.

In contrast, the CDK2 p-SIFt is more similar to the ATPg p-SIFt as can be observed in the CDK2-ATP difference profile. Unlike p38, in CDK2 the Phe "gatekeeper" residue blocks access to the hydrophobic pocket. As a result, many of the residues accessible to CDK2 inhibitors will be those that also interact with ATP. In fact, of the conserved residues observed in the CDK2 p-SIFt, there are none that are not also conserved in the ATPg p-SIFt. The main positive difference regions of the CDK2-ATP difference profile, corresponding to intermediate level conserved interactions in the CDK2 p-SIFt that occur with low frequency in the ATP p-SIFt are colored white in Figure 3d.

Unlike contacts with the hydrophobic pocket, several interactions conserved in the p38 cluster are common to CDK2, as well as other non-ATP inhibitors, and are colored red in Figure 3d. Finally, several interactions are conserved for ATPg and are observed with relatively low frequency for CDK2 and p38. These ATPg specific contacts are colored yellow in Figure 3d and involve residues at positions 50−55, which interact with the ribose and phosphate moieties of ATP, and with residues at positions 168, 170, and 171, in the vicinity of the catalytic loop.

Canonical Interactions. Approximately 20% of the contact interactions are conserved in each of the ATPg, CDK2, and p38 p-SIFts as well as over the 93 structures as a whole. These are denoted in Table 2 by the highlighted annotations and comprise a canonical set of interactions that are evidently fundamental for kinase binding at the ATP site. Further analysis of the full length SIFts revealed that among this set are interactions with residues at positions 121 and 123, which are involved in hydrogen bonding to the adenine moiety of ATP, the "gatekeeper" residue, position 57 in the glycine rich loop, position 70 that for ATP involves hydrophobic interactions between adenine and β3, and position 72 involving the ATP phosphates interacting with β3. The residues involved in the canonical set of interactions are colored green in Figure 3d.

The canonical interactions comprise an essential kinase-binding signature for compounds targeting the ATP binding site. Although as noted in Table 2, additional conserved interactions exist for the ATPg, p38, and CDK2 clusters, the canonical interactions are common to all inhibitors and may be used as a basic kinase like binding filter in VS.

3. Subclustering of p38 SIFts. Hierarchical clustering of the SIFts computed from the 93 kinase X-ray structures resulted in the identification of the p38 and CDK2 clusters because they represent two fundamentally different sets of small molecule inhibitors in terms of interactions with the ATP binding site. However, the SIFts within each cluster are not homogeneous. In particular, the p38 cluster reveals interesting details on the relationship between interaction patterns and inhibitor selectivity.

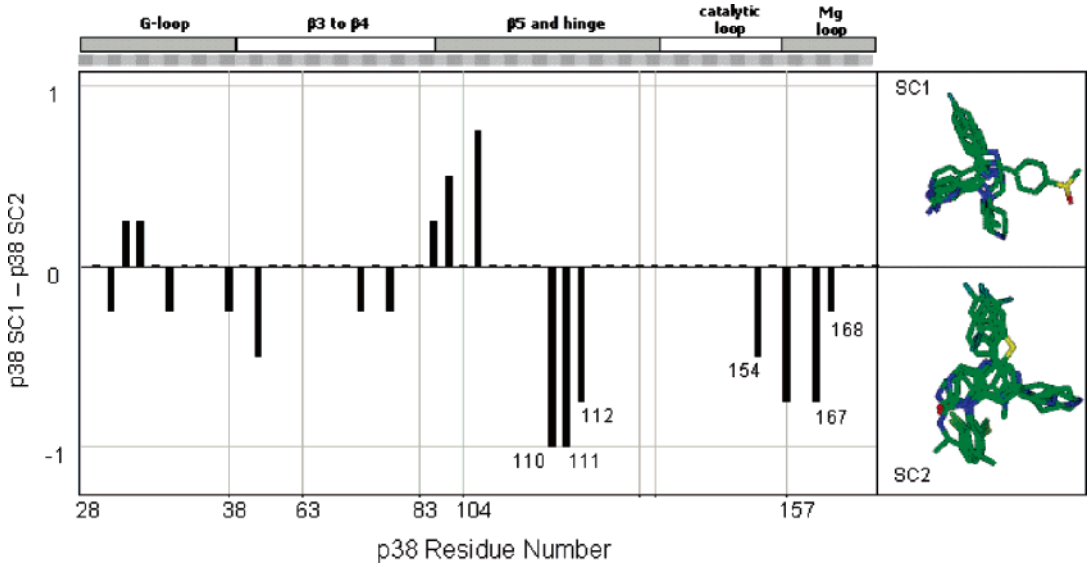

Figure 5. Difference profile plot derived from the clustering of p38 inhibitors. Subcluster 1 (SC1) corresponds to the well-known pyridinyl imidazole class of inhibitors whereas subcluster 2 (SC2) contains several more recently reported inhibitors. Residues showing key interaction differences between the two classes of inhibitors, as discussed in the text, are labeled on the plot.

Clustering of the nine structures comprising the p38 cluster identifies three distinct SIFt subclusters representing two distinct classes of inhibitors (shown in Figure 5) displaying overall similar yet unique binding signatures at the ATP binding site. The contact only difference profile plotted clearly shows that while the two structurally different classes of inhibitors share a common set of interactions with the kinase, each class has marked regions where the p-SIFts differ. In particular, subcluster 2 has additional contacts with β5, the hinge region, Mg loop, and catalytic regions of the kinase, clearly visible in the difference profile. Of particular interest are members of subcluster 2 that are more potent inhibitors of p38 and have been reported to exhibit improved selectivity against Erk and Jnk.[51,52] Scapin et al. have rationalized that the improved selectivity of the subcluster 2 inhibitors is due to a peptide bond flip between Met109 and Gly110 in the hinge induced by the inhibitors and accommodated by the small side chains in p38 relative to Erk and Jnk. The difference profile clearly shows the resulting additional interactions at positions 110, 111, and 112 (p38 numbering) exploited by the subcluster 2 inhibitors. In addition to the interactions reported previously as the structural basis for improved subcluster 2 inhibitor potency and selectivity, the p-SIFts also reveal additional contacts with the Mg-loop and catalytic loop regions of p38.

Results from the analysis of the p38 cluster illustrate the power of the p-SIFt approach, namely, the ability to quantify the similarities and differences in the interaction patterns of inhibitors to a given target. Moreover, the ability to derive p-SIFts and difference profiles that quantify key conserved interactions can form the basis for inferring the structural basis for inhibitor potency and selectivity. The detailed binding signature information encoded in the p-SIFts make them ideal filters for screening virtual libraries, discussed below in Section 4.

4. Virtual Screening. In the previous sections we have identified clear conservation patterns of interactions observed for ATP, p38, and CDK2 clusters as well as a canonical set of conserved interactions common to all ligands bound to kinases at the ATP binding site. We now apply these binding signatures to VS for protein kinase inhibitors.

The success of VS methodologies is typically cast in terms of enrichment studies designed to measure the percentage of known actives identified as a function of the fraction of the database screened. Often, the results of these studies indicate that the performance of scoring functions is target specific, for example, leading to significant enrichment of actives for docking against the estrogen receptor but performing poorly against kinase targets.[53] Unfortunately, knowledge of the optimal scoring function to apply in a virtual screen against a novel target is not available a priori. As a result, it is often necessary to undertake lengthy validation studies to select a suitable scoring function, or alternately, construct a customized scoring scheme optimized for the target of interest. These problems are compounded when VS is carried out against multiple targets.

Some of these difficulties can be addressed by applying p-SIFts to the ranking and filtering of VS results. p-SIFts are in essence target specific molecular filters encoding binding signatures that are consistent with a particular target specific group of known active inhibitors. Moreover, by comparing the SIFt for each docked solution with a kinase specific, or binding mode specific, p-SIFt, each p-SIFt is in effect a target specific scoring function. In this section we will demonstrate how the p-SIFt can be applied in a VS workflow that can be tailored to a specific target without having to rely on the ambiguities of energy-based scoring.

To this end, we have tested the performance of p-SIFt-based scoring in a typical database enrichment application using p38 and CDK2 as targets. In addition, the degree to which the ATP, CDK2, and p38 p-SIFts are selective toward observed kinase inhibitor binding modes was also assessed. Finally, for the generation of enrichment curves and selectivity assessment tests, full-length p-SIFts derived from 7-bit SIFts were used.

Database Enrichment. A database containing known inhibitors of both p38 and CDK2, in a background of

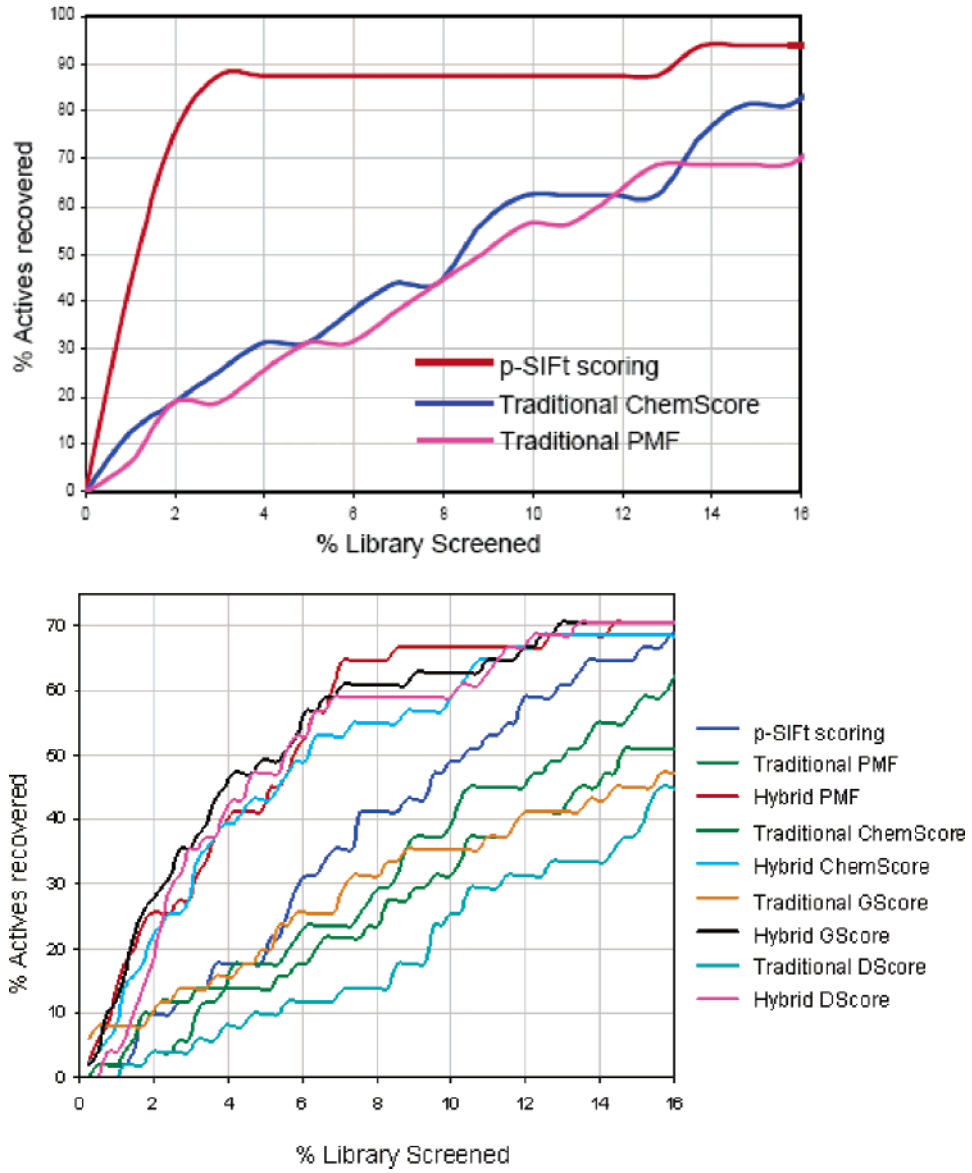

Figure 6. (a) Enrichment curves obtained for VS against p38. The enrichment curves were derived using the scoring methods described in the Methods section. The ChemScore and PMF curves were obtained using the traditional scoring scheme, using the ChemScore and PMF scoring functions, respectively, for both final pose selection and ligand ranking. Other scoring functions performed similarly under the traditional scoring scheme. (b) Enrichment curves obtained for VS against CDK2. The enrichment curves were derived using the scoring methods described in the Methods section. The traditional PMF, ChemScore, GScore, and DScore curves were obtained using the traditional scoring scheme, using the indicated scoring function, respectively, for both final pose selection and ligand ranking. The hybrid PMF, ChemScore, Gscore, and DScore curves were obtained by applying the hybrid scoring scheme using the indicated function for both final pose selection and ligand ranking.

1000 diverse commercially available compounds was docked against the X-ray structures of CDK2 (PDB code 1di8) and p38 (PDB code 1a9u). The ability of our p-SIFt VS protocol to identify known actives was quantified by computing enrichment curves plotting the percentage of actives recovered as a function of the percentage of the database screened.

p38. In this section we present database enrichments for p38 obtained from VS using the traditional and p-SIFt scoring methods detailed in the Methods section. Enrichment curves and cumulative enrichment factors for p38 are presented in Figure 6a comparing the traditional and p-SIFt scoring approaches.

From Figure 6a it is clear that the enrichment obtained by applying p-SIFt scoring provided markedly superior results over those obtained using traditional scoring using the Chemscore and PMF functions. Moreover, there is little difference between these functions over the first 15% of the database. In contrast, p-SIFt scoring performs close to the ideal enrichment curve over the first 2% of the database, meaning that 14 of the 16 known p38 actives were in the top 20 ranked ligands. Upon examination of the docking poses, it was discovered that for two inhibitors correct poses were never generated in the initial pose pool. The p-SIFt scoring method requires a pose having a correct docked binding mode to generate a high Z_{p38} value, unlike traditional scoring which can generate high scores even for poses that bind incorrectly. Generating enrichments for the right reasons is a built-in advantage of the p-SIFt scoring approach. For p38, the hybrid scoring scheme

was found to offer no improvement in enrichment over that obtained from using p-SIFt scoring.

CDK2. Enrichment curves were derived using traditional scoring, p-SIFt scoring, and hybrid scoring, are presented in Figure 6b for docking against CDK2. A striking feature of these curves is that all of the hybrid scoring scheme variants performed better than the traditional and p-SIFt schemes irrespective of what scoring function was used for pose selection and ranking. It would appear that once the majority of the incorrect poses that contribute to false positive scores are filtered out, the differences between scoring functions visible in the results using these functions alone (traditional scoring) is factored out. Enrichments obtained using p-SIFt scoring are comparable to traditional scoring up to 6% of the database screened and significantly better at higher levels.

Attaining database enrichments for CDK2 comparable to those obtained for p38 is a considerably more challenging task for VS. The large gatekeeper residue in CDK2 restricts the number of residues accessible in the ATP binding site. The p-SIFt for CDK2 samples fewer residues compared to p38 and conserved interactions are distributed over a relatively small spatial region. As a result, in the CDK2 there are fewer constraints to generate ligand placements and it is therefore easier to generate poses that satisfy conserved interactions in CDK2 compared to p38 where the residues of the hydrophobic pocket are accessible. In effect, the CDK2 p-SIFt is less selective against false poses as evidenced by the superior performance of p-SIFt scoring for p38 versus CDK2.

Selectivity Assessment. The difference profiles presented in Section 2 exhibit clear regions where ATPg, CDK2, and p38 inhibitors bind to kinases in unique ways. These observations suggest that p-SIFts can be used to model the selectivity of inhibitors based on the types of interactions they are able to satisfy when binding to the kinase. To validate the use of p-SIFts as selectivity filters, we have carried out self-recognition experiments using the set of 93 X-ray structures as a test data set. For this purpose, ATPg, CDK2, and p38 validation p-SIFts were derived. For CDK2 and ATPg ~50% of the structures in each group were selected randomly to derive the p-SIFt, whereas for p38, only the SIFts corresponding to subcluster 1 were used. The remaining ATPg, CDK2, and p38 structures were not used to derive the validation p-SIFts. For each p-SIFt, Z_{target} values were then computed against each of the 93 kinase structures in order to assess the ability of p-SIFts to recognize members of their own group. For the p-SIFts to serve as effective molecular filters, the p38 p-SIFt needs to generate statistically significantly higher Z_{p38} against the p38 cluster X-ray structures relative to the remaining structures, whereas the CDK2 and ATPg p-SIFts should perform well against the CDK2 and ATPg structures, respectively.

It is clear from Figure 7 that for each p-SIFt, the distribution for the corresponding set of target structures is shifted toward higher Z-scores. Considering ATPg first, the top scoring 26% of the total 93 structures are ATPg cluster members, making up approximately 65% of all of the ATPg. The remaining ATPg structures fall into a region of the distribution that overlaps with the distribution for CDK2, p38, and the remaining structures. Interestingly, the overlap in the distributions

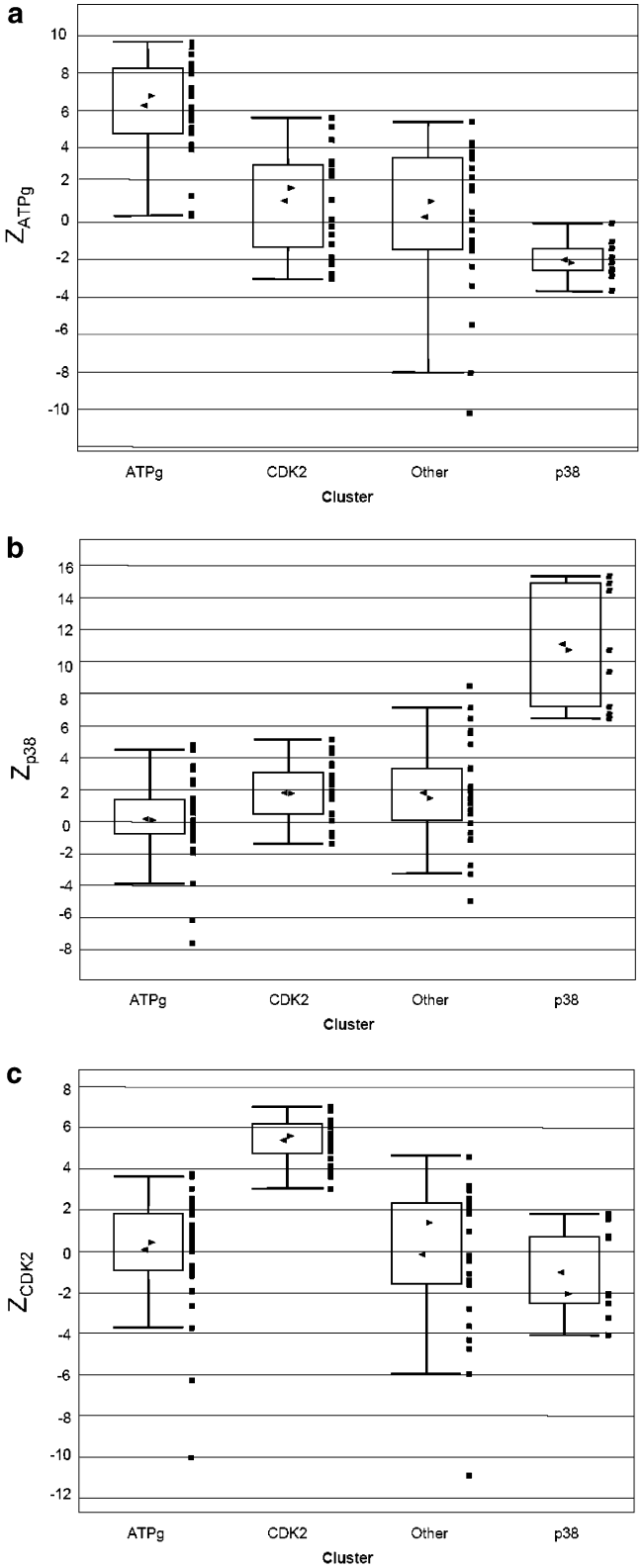

Figure 7. Box plots of Z_{target} distributions obtained for the ATPg, p38, and CDK2 cluster subsets described in the text are shown for all kinases in the 93 X-ray structure set in panels a−c, respectively. The right and left arrows indicate the mean and the median, respectively, of the distribution; the vertical error bars delineate the upper and lower adjacent values of the data; the square markers represent individual data points. The box outlines the second and third quartiles of the distribution.

may be rationalized in terms of the p-SIFt similarities discussed in Section 3. The similarity between the ATPg and CDK2 p-SIFts, discussed previously, has the consequence that 90% of the CDK2 structures overlap in Z with the lowest scoring 35% of the ATPg. This overlap exists primarily because the ATPg p-SIFt is in essence derived from a subset of the interactions sampled by CDK2. However, the differences between the ATP and CDK2 interaction patterns are captured in the CDK2 p-SIFt. Consequently, the highest segment in the distribution shown in Figure 7c contains 19 of 20 CDK2 structures and overlaps with only 2 ATPg structures.

The greatest separation in Z-score distributions is obtained for p38, due primarily from p-SIFt features reflecting conserved residues in the hydrophobic pocket of the ATP binding site. The p38 structures fall into two groups in the distribution, dividing neatly between the highest scoring subcluster 1 structures, used to derive the p-SIFt, and slightly lower scoring subcluster 2 examples. The latter set overlaps in Z_{p38} with 1qpe, 1pme, and 3erk. Both 1pme and 3erk are examples of complexes between pyridinyl imidazole compounds complexed with variants of Erk2, whereas 1qpe is a structure of PP2 complexed with Lck. As in p38, the "gatekeeper residue" in Lck is also Thr, and the two kinases have relatively similar ATP binding sites. These examples highlight the fact that the p-SIFts are able to capture similarities in interaction patterns arising, on one hand from ligand similarity (1pme and 3erk), and on the other, from binding site similarity. Finally, we find that that using a multistructure p-SIFt rather than a single structure (1a9u) yields improved separation between Z-score distributions.

Discussion and Conclusion

This paper introduces interaction profiling (p-SIFt) as a new and powerful approach to understand small molecule protein interactions. Sequence profile approaches have previously been shown to successfully detect remote protein sequence and structural relationships. Our results are the first demonstration, to our knowledge, that this approach seems well suited to analysis of protein−inhibitor complexes.

The implementation of the p-SIFt methodology and underlying SIFts is very flexible. The SIFts themselves can be constructed from any number of interaction features or bits, whereas the p-SIFts can be derived to capture very general interaction patterns, for example, a canonical kinase binding signature, or rather, capture a very focused set of interactions, as in the example of the p38 subcluster difference profile. We used simple averaging of the bit values to construct the profiles, although more sophisticated weighting schemes, as described extensively in the literature,[54] can also be used in order to improve the sensitivity of the p-SIFt. The weighting scheme may be useful to reduce the contribution of highly correlated bits in the profile or to add weight to bits that are most correlated with inhibitor activity. In addition to the Tanimoto coefficient, several other well-known similarity metrics can be applied to compare the bit-strings, such as the cosine coefficient and measures of the Euclidean distance.[44−47]

The results from the profile analysis of ATP, p38, and CDK2 show that these profiles can be useful to quickly define similarities and differences between groups of small molecule inhibitors and these can be applied to VS. We believe that these examples are generally applicable to other important protein−drug targets. It is apparent that in order to be competitive with ATP, small molecule inhibitors must satisfy a set of canonical interactions, common to those made by the adenine moiety of ATP, along the hinge and $\beta3-\beta5$ regions of the kinase. In contrast, conserved interactions involving the ATP-phosphates and ribose ring are not a requirement for small molecule inhibition. With few exceptions, small molecule inhibitors and ATP share the same "anchoring" interactions at the hinge region of the binding site, but not necessarily the interactions key to metal binding and phosphorylation. Our profile−profile comparisons also reveal that, unlike ATP, on average, small molecules interact with the kinase in the region of the hydrophobic pocket when accessible and extend further along the hinge toward the solvent interface of the ATP binding site.

We find that when comparing the overall conserved and variable interactions observed in different kinase subgroups, the coarse-grained contact-only p-SIFts provide a clear and uncluttered analysis. However, for detailed analyses and in VS applications, the full-featured interaction p-SIFt is most useful.

Virtual screening, in general, suffers from a signal-to-noise problem: the true ligand binding mode is more often than not buried in a background of high scoring incorrect ligand placements.[22] Because of this high false positive problem inherent in VS, the ability to filter out "bogus" poses is critical.[27] Our results show that the Z-score computed against a p-SIFt is a sensitive measure, able to separate ligands that are binding to a receptor with a required interaction pattern from high scoring VS poses that are binding inconsistently with known ligands to the desired target. Of course, p-SIFt filtering may also miss out novel inhibitor binding modes (false negatives). However, the p-SIFt approach may be also used to identify novel binding modes by retaining poses that are most dissimilar to the p-SIFt derived from known binding modes. This approach could be extended in order to select a diverse set of binding modes spanning a range of similarity with the p-SIFt.

Cross-docking experiments present a particularly challenging problem for VS. Docking tends to generate many bogus poses per ligand, creating the daunting task of detecting the correct poses from many incorrect ligand placements. To further complicate matters, in cross-docking some of the incorrect poses will correspond to inhibitors of another kinase and therefore contain kinase-binding motifs, increasing the probability that they will score as false-positive hits in the VS. However, our results demonstrate that the p-SIFt-based scoring scheme can serve as a flexible and transferable target specific approach to VS lead discovery.

Given the rapid growth in the number of available X-ray structures, it should be possible to eventually construct and screen against a virtual selectivity panel of p-SIFts for multiple drug targets in much the same way that inhibitors are routinely tested against a panel of in vitro inhibition assays. The resulting virtual selectivity panel could be precomputed for ligands in virtual libraries, thus providing an annotation that

could be mined when selective inhibitors to any target are desired. In addition, predicted cross-reactivity to a given drug target could be an effective starting point for lead discovery for novel targets, an approach that has been demonstrated to be fruitful for protein kinases.

Acknowledgment. The authors would like to thank Christian Lemmen, Marcus Gastreich, and Rajiah Denny for their help with the docking experiments. We also thank Rainer Fuchs, Herman van Vlijmen, Alexey Lugovskoy, and Jennifer Campbell for helpful comments on the manuscript.

Appendix

The abbreviations of common amino acids are in accordance with the recommendations of IUPAC. Additional abbreviations: ATP, adenosine triphosphate; ATPg, adenosine triphosphate group; CDK2, cyclin dependent kinase 2; Chk1, checkpoint kinase 1; EFCP, extended connectivity fingerprint; Erk, extracellular signal-regultaed kinase; Jnk, c-Jun N-terminal kinase; NMR, nuclear magnetic resonance; p38, p38 mitogen-activated protein kinase; PKA, protein kinase A; PDB, protein databank; p-SIFt, profile structural interaction fingerprint; SBDD, structure-based drug design; SC1, subcluster 1; SC2, subcluster2; SIFt, structural interaction fingerprint; TGFβ-RI, transforming growth factor beta receptor I; VS, virtual screening.

Supporting Information Available: A table listing the 93 X-ray structures and corresponding ligands used in this paper is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Alvarez, J. C. High-throughput docking as a source of novel drug leads. *Curr. Opin. Chem. Biol.* **2004**, DOI 10.1016/j.cbpa.2004.05.001.
(2) Perola, E.; Walters, W. P.; Charifson, P. S. A detailed comparison of current docking and scoring methods on systems of pharmaceutical relevance. *Proteins* **2004**, *56*, 235−249.
(3) Muegge, I.; Enyedy, I. J. Virtual screening for kinase targets. *Curr. Med. Chem*. **2004**, *11*, 693−707.
(4) Lyne, P. D. Structure-based virtual screening: an overview. *Drug. Discovery Today* **2002**, *7*, 1047−1055.
(5) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882−894.
(6) Chalk, A. J.; Worth, C. L.; Overington, J. P.; Chan, A. W. PDBLIG: Classification of Small Molecular Protein Binding in the Protein Data Bank. *J. Med. Chem.* **2004**, *47*, 3807−3816.
(7) 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.
(8) Gribskov, M.; McLachlan, A. D.; Eisenberg, D. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. U.S.A.* **1987**, *84*, 4355−4358.
(9) Gribskov, M.; Luthy, R.; Eisenberg, D. Profile analysis. *Methods Enzymol.* **1990**, *183*, 146−159.
(10) Wang, G.; Dunbrack, R. L., Jr. Scoring profile-to-profile sequence alignments. *Protein Sci.* **2004**, *13*, 1612−1626.
(11) Mehta, P. K.; Argos, P.; Barbour, A. D.; Christen, P. Recognizing very distant sequence relationships among proteins by family profile analysis. *Proteins* **1999**, *35*, 387−400.
(12) Rice, D. W.; Eisenberg, D. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* **1997**, *267*, 1026−1038.
(13) Koonin, E. V.; Wolf, Y. I.; Aravind, L. Protein fold recognition using sequence profiles and its application in structural genomics. *Adv. Protein Chem.* **2000**, *54*, 245−275.
(14) Stahl, M.; Rarey, M. Detailed analysis of scoring functions for virtual screening. *J. Med. Chem.* **2001**, *44*, 1035−1042.
(15) McGovern, S. L.; Caselli, E.; Grigorieff, N.; Shoichet, B. K. A common mechanism underlying promiscuous inhibitors from virtual and high-throughput screening. *J. Med. Chem.* **2002**, *45*, 1712−1722.
(16) 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.
(17) Kontoyianni, M.; McClellan, L. M.; Sokol, G. S. Evaluation of docking performance: comparative data on docking algorithms. *J. Med. Chem.* **2004**, *47*, 558−565.
(18) Schulz-Gasch, T.; Stahl, M. Binding site characteristics in structure-based virtual screening: evaluation of current docking tools. *J. Mol. Model.* **2003**, *9*, 47−57.
(19) Jansen, J.; Martin, E. Target-biased scoring approaches and expert systems in structure based virtual screening. *Curr. Opin. Chem. Biol.* **2004**, in press.
(20) Wang, R.; Lu, Y.; Wang, S. Comparative evaluation of 11 scoring functions for molecular docking. *J. Med. Chem.* **2003**, *46*, 2287−2303.
(21) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.
(22) Lyne, P. D.; Kenny, P. W.; Cosgrove, D. A.; Deng, C.; Zabludoff, S.; Wendoloski, J. J.; Ashwell, S. Identification of compounds with nanomolar binding affinity for checkpoint kinase-1 using knowledge-based virtual screening. *J. Med. Chem.* **2004**, *47*, 1962−1968.
(23) Paul, N.; Kellenberger, E.; Bret, G.; Muller, P.; Rognan, D. Recovering the true targets of specific ligands by virtual screening of the protein data bank. *Proteins* **2004**, *54*, 671−680.
(24) Jacoby, E.; Schuffenhauer, A.; Floersheim, P. Chemogenomics knowledge-based strategies in drug discovery. *Drug News Perspect.* **2003**, *16*, 93−102.
(25) Klon, A. E.; Glick, M.; Thoma, M.; Acklin, P.; Davies, J. W. Finding more needles in the haystack: A simple and efficient method for improving high-throughput docking results. *J. Med. Chem.* **2004**, *47*, 2743−2749.
(26) Fradera, X.; Mestres, J. Guided docking approaches to structure-based design and screening. *Curr. Top. Med. Chem.* **2004**, *4*, 687−700.
(27) Claussen, H.; Gastreich, M.; Apelt, V.; Greene, J.; Hindle, S. A.; Lemmen, C. The FlexX Database Docking Environment − Rational Exctraction of Receptor Based Pharmacophores. *Curr. Drug Discovery Technol.* **2004**, *1*, 49−60.
(28) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209−225.
(29) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.
(30) Vangrevelinghe, E.; Zimmermann, K.; Schoepfer, J.; Portmann, R.; Fabbro, D.; Furet, P. Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking. *J. Med. Chem.* **2003**, *46*, 2656−2662.
(31) Kroemer, R. T.; Vulpetti, A.; McDonald, J. J.; Rohrer, D. C.; Trosset, J. Y.; Giordanetto, F.; Cotesta, S.; McMartin, C.; Kihlen, M.; Stouten, P. F. Assessment of docking poses: interactions-based accuracy classification (IBAC) versus crystal structure deviations. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 871−881.
(32) Singh, J.; Chuaqui, C. E.; Boriack-Sjodin, P. A.; Lee, W. C.; Pontz, 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). *Bioorg. Med. Chem. Lett.* **2003**, *13*, 4355−4359.
(33) Cohen, P. Protein kinases—the major drug targets of the twenty-first century? *Nat. Rev. Drug. Discovery* **2002**, *1*, 309−315.
(34) ter Haar, E.; Walters, W. P.; Pazhanisamy, S.; Taslimi, P.; Pierce, A. C.; Bemis, G. W.; Salituro, F. G.; Harbeson, S. L. Kinase chemogenomics: targeting the human kinome for target validation and drug discovery. *Mini Rev. Med. Chem.* **2004**, *4*, 235−253.
(35) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298*, 1912−1934.
(36) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243−257.
(37) Adams, J.; Lee, D. Recent progress towards the identification of selective inhibitors of serine/threonine protein kinases. *Curr. Opin. Drug Discovery Dev.* **1999**, *2*, 96−109.
(38) PipelinePilot; Scitegic Inc., San Diego, California, version 3.0.1.0.
(39) Omega; Openeye Scientific Software Inc., Santa Fe, New Mexico, version 1.8.
(40) 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.

(41) Kramer, B.; Rarey, M.; Lengauer, T. Evaluation of the FLEXX incremental construction algorithm for protein−ligand docking. *Proteins* **1999**, *37*, 228−241.
(42) Sybyl; Tripos, Inc., St. Louis, MO, version 6.9.1.
(43) Willett, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(44) Chen, X.; Reynolds, C. H. Performance of similarity measures in 2D fragment-based similarity searching: comparison of structural descriptors and similarity coefficients. *J Chem. Inf. Comput. Sci.* **2002**, *42*, 1407−1414.
(45) Raymond, J. W.; Blankley, C. J.; Willett, P. Comparison of chemical clustering methods using graph- and fingerprint-based similarity measures. *J. Mol. Graph. Model.* **2003**, *21*, 421−433.
(46) Willett, P. Similarity-based approaches to virtual screening. *Biochem. Soc. Trans.* **2003**, *31*, 603−606.
(47) Willett, P. Evaluation of molecular similarity and molecular diversity methods using biological activity data. *Methods Mol. Biol.* **2004**, *275*, 51−64.
(48) Dubes, R.; Jain, A. K. Clustering methodologies in exploratory analysis. *Adv. Comput.* **1980**, *19*, 113−228.
(49) CScore; Tripos, Inc., St. Louis, MO, version in Sybyl 6.9.1.
(50) See structures with PDB codes: 1atp, 1phk, 2phk, and 1ql6.
(51) Fitzgerald, C. E.; Patel, S. B.; Becker, J. W.; Cameron, P. M.; Zaller, D.; Pikounis, V. B.; O'Keefe, S. J.; Scapin, G. Structural basis for p38alpha MAP kinase quinazolinone and pyridol-pyrimidine inhibitor specificity. *Nat. Struct. Biol.* **2003**, *10*, 764−769.
(52) Scapin, G. Structural biology in drug design: selective protein kinase inhibitors. *Drug Discovery Today* **2002**, *7*, 601−611.
(53) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.
(54) Vingron, M.; Sibbald, P. R. Weighting in sequence space: a comparison of methods in terms of generalized sequences. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 8777−8781.
(55) Eldridge, M.; Murray, C. W.; Auton, T. A.; Paolini, G. V.; Lee, R. P. Empirical scoring functions: I. The development of a fast emprical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425−445.
(56) 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.
(57) Muegge, I.; Martin, Y. C. A general and fast scoring function for protein−ligand interactions: a simplified potential approach. *J. Med. Chem.* **1999**, *42*, 791−804.
(58) Meng, C.; Shoichet, B. K.; Kuntz, I. D. Automated docking with grid-based energy evaluation. *J. Comput. Chem*. **1992**, *13*, 505−524.
(59) Charifson, P. S.; Corkery, J. J.; Murcko, M. A.; Walters, W. P. Consensus scoring: A method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem.* **1999**, *42*, 5100−5109.

JM049312T