Knowledge-Based Design of Target-Focused Libraries Using Protein−Ligand Interaction Constraints

Zhan Deng, Claudio Chuaqui, and Juswinder Singh*

Computational Drug Design Group, Biogen Idec, Inc. 12 Cambridge Center, Cambridge, Massachusetts 02142

Received April 22, 2005

Here we present a new strategy for designing and filtering potentially massive combinatorial libraries using structural information of a binding site. We have developed a variation of the structural interaction fingerprint (SIFt) named r-SIFt, which incorporates the binding interactions of variable fragments in a combinatorial library. This method takes into account the 3D structure of the active site of the target molecule and translates desirable ligand−target binding interactions into library filtering constraints. We show using the MAP kinase p38 as a test case that we can efficiently analyze and classify compounds on the basis of their abilities to interact with the target in the desired binding mode. On the basis of these classifications, decision tree models were generated using the molecular descriptors of the compounds as predictor variables. Our results suggest that r-SIFt coupled with the classification models should be a valuable tool for structure-based focusing of combinatorial chemical libraries.

Introduction

The past decade has witnessed significant advances in combinatorial chemistry. With the discovery and availability of more reagents and reaction schemes, as well as the advances of chemical synthesis methods, the number of compounds that are synthetically feasible is dauntingly massive.[1] In recent years, a great amount of research effort has been focused on how to design smaller libraries that are tailored to specific drug targets or gene-families, instead of generating large, diverse, and general purpose libraries, with an aim to make the lead discovery and optimization processes more efficient and cost-effective.[2] Several strategies and techniques in designing subsets of combinatorial libraries have been previously discussed.[3−7]

Parallel with the advances in chemical library synthesis, high-throughput X-ray crystallography and NMR techniques are becoming more and more sophisticated, generating large numbers of experimental 3D structures of drug target molecules. This structural information provides great insights into the activity, mechanism and regulation of drug target molecules, and it provides the basis of structure-based drug design.[8,9]

Some techniques have been available for leveraging the structural information of the target molecules into library design and filtering. One example is the 3D pharmacophore model. If a collection of known active molecules are available, abstract 3D pharmacophore models can be generated from these compounds by extracting the common spatial arrangement of pharmacophoric features. These models can be applied to filter a large library and identify compounds that also satisfy the pharmaophore.[10] An alternative and more directed approach is termed structure-based focusing (SBF). This method enables combinatorial chemical libraries to be tailored to binding sites using specific interaction constraints.[11] However, the definition of which constraints to use in structure-based focusing is somewhat ad-hoc and not systematic.

We have recently developed structural interaction fingerprint (SIFt), a novel method for efficiently representing, visualizing, and analyzing massive amounts of structures. SIFt has been proven to be useful in facilitating post-docking analysis, virtual screening, and database mining of structural data.[11,12] Key to the SIFt method is the generation of structural interaction fingerprints, 1D binary bit-strings that represent important target−ligand binding interactions, thus making the structures amenable to easy mathematical manipulation and comparison. We have shown that by combining SIFt and other conventional scoring functions, we can achieve much better confidence in reproducing the true binding modes of the compounds and thereby obtain improved library enrichments from virtual screening.[11,12]

In silico virtual screening and computer-aided drug design have become increasingly important in drug discovery.[13,14] How to intelligently leverage the 3D structural information of target molecules and use it in designing target-focused libraries is of great interest in the field. In this paper, we propose a new strategy for designing and filtering target-specific combinatorial libraries. To this purpose, we designed r-SIFt, a variation of the original SIFt, that incorporates the binding information pertinent to different variable fragments of a combinatorial library into the fingerprint. The “r” in r-SIFt stands for the various R groups of a combinatorial library. In r-SIFt, the binary bits represent whether a particular R group or the core fragment of a compound is interacting with a particular residue of the target molecule. We show that this variation of SIFt provides a way to directly and conveniently visualize how different fragments of the ligands are placed in the active site and that r-SIFt is very sensitive and effective in separating different binding modes, therefore making it especially useful for analyzing and organizing the virtual screening results of a combinatorial library.

We applied the r-SIFt method to classify compounds on the basis of their abilities to interact with the target in a desired binding mode. On the basis of these classifications, we built machine learning models and demonstrated that these predictive models can effectively enrich large libraries to generate a subset of compounds that are more likely to adopt the same desirable binding mode. We have tested this strategy on several different combinatorial libraries of MAP kinase p38 inhibitors. The results demonstrate that the predictive models based on r-SIFt clas-

* To whom correspondence should be addressed. Tel: (617) 679-2027. Fax: (617) 679-3635. E-mail: singhjus@yahoo.com.

10.1021/jm050381x CCC: $33.50
Published on Web 12/17/2005

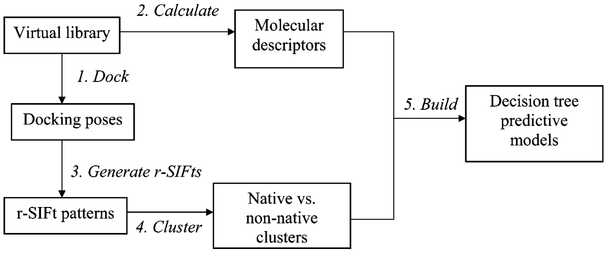

Figure 1. An overview of the major computational steps in r-SIFt analysis.

sification can be used as effective filters to eliminate poor binders from a large combinatorial library in various stages of lead discovery and optimization, thus generating smaller, more efficient, target-focused libraries for subsequent screening.

Materials and Methods

Our r-SIFt analysis consists of the following five major computational steps: library enumeration and docking, calculation of 2D molecular descriptors of the variable R groups, construction of r-SIFt patterns, clustering and classification of r-SIFt patterns, and generation of predictive models. The relationship of these five major components is summarized in Figure 1.

1. Virtual Library Enumeration and Docking. We generated several virtual chemical libraries and ensembles of docking poses for our analysis. A crystal structure of MAP kinase p38 (PDB accession code 1ouk)[15] was used as the target molecule in all of our virtual screening experiments. The first set of docking poses was used to demonstrate the ability of r-SIFt to efficiently differentiate and visualize different binding modes. The pyridinyl imidzole inhibitor cocrystallized with p38 in the 1ouk structure (Figure 3). **1**, which has been identified by Merck to be a very selective and potent p38 inhibitor, was docked onto the original target molecule. We retained 150 poses with the highest Cscores[22] for subsequent analysis. The docking experiments were carried out with FlexX[16] in Sybyl.[17] The ligand binding site was defined using a cutoff radius of 10 Å from the 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 during the docking. Five different scoring functions, including Fscore,[16] ChemScore,[18] Gscore,[19] PMF Score,[20] and Dscore,[21] were used as voting scores in the Cscore[22] utility in Sybyl. Figure 2a shows these 150 poses of **1** generated from the docking experiment. They display a variety of binding modes at the active site of the target protein.

To compare and contrast the r-SIFt patterns of different compound structures, we also performed docking experiments using five chemically distinct compounds (Figure 3), using the same FlexX docking procedure. These five compounds are **1**, discussed above (PDB code 1ouk); **2** (SB203580), a well-known pyridinyl imidazole p38 inhibitor;[23] **3** (SKF-86002), a compound discovered by SmithKline Beecham;[24] **4** (2-(2,6-dichlorobenzyl)-5-(4-fluorophenyl)-6-pyridin-4-methyl-5*H*-pyrimidin-4-one), a compound first reported by Amgen;[24] and **5** (2-[1,3]dithietan-2-ylidene-2-pyridin-4-yl-1-(4-trifluoromethoxyphenyl)ethanone), a molecule that exhibits no p38 inhibition activity.[25] Except for **5**, the other four compounds are all known to be potent p38 inhibitors.[15,24] At the time this paper was being prepared, we were not aware of the cocrystal structure of **3**, **4**, or **5** with p38 in public databases. Figure 3 shows the 2D chemical structures of these compounds, including our definitions of their cores and variable R groups. For all docking experiments, top 10 poses of each compound with the highest Cscores were retained for subsequent analysis.

In addition, to test our r-SIFt method on combinatorial libraries, we enumerated five different libraries using three distinct p38 inhibitors as template scaffolds, **1**, **3**, and **4**, varying only one R group at a time in each library. A common set of reagents consisting of about 10,000 commercially available aryl bromides[26] was used as R groups in the enumeration of these libraries. Three libraries (1-R1, 3-R1, and 4-R1) were enumerated by varying the R1 group of the three templates, respectively. The fourth and the fifth libraries (1-R2 and 1-R3) were based on **1**, varying R2 and R3 groups, respectively (Table 1). According to the cocrystal structures of 1ouk and other similar inhibitors (PDB codes 1a9u, 1bl6, 1bl7, 1bmk, 1ove, etc.),[15,23] in the "native binding mode", the R1 groups are expected to interact with the hydrophobic pocket[27] of p38. The R2 portion of **1**, on the other hand, is positioned in the vicinity of the adenine binding site in the hinge region, whereas the R3 group interacts with the phosphate binding region (P-loop).[15,27] Library enumeration processes were carried out using Pipeline Pilot.[28] All the reaction products were prefiltered by removing salts, isotopes, and inorganic compounds as well as molecules with molecular weight less than 400. From the remaining library, a subset of molecules (maximum number 2500) with maximal chemical diversity was sampled for further analysis. The total number of selected compounds of each library is as follows: 1-R1, 2208; 1-R2, 2450; 1-R3, 2000; 3-R1, 2442; 4-R1, 1750.

These five libraries were docked onto the p38 target molecule (1ouk), using the same docking procedure as previously described. Ten poses with the best Cscores for each molecule were saved for further r-SIFt analysis (see sections 3 and 4). Reassuringly, the docking experiments were able to reproduce the native cocrystal structure of **1**, with an rmsd (for heavy atoms) less than 0.6 Å, confirming the validity of the docking procedure.

2. Calculation of 2D Descriptors. 2D molecular descriptors of the R group monomers (after substituting the bromide with a hydrogen atom) were calculated using Pipeline Pilot.[28] To make the method more amenable to huge libraries, we decided to omit the time-consuming calculation of 3D descriptors. A total of 37 2D descriptors were generated.

The molecular descriptors set was further processed by removing variables with little or no variance across the whole library. In addition, descriptors with high redundancy and multicollinearity were removed. This cleaning step was performed using the unsupervised forward selection (UFS) algorithm[27] with the stopping criteria of $R_{\max}^2$ (i.e., the squared multiple correlation coefficient, SMCC) cutoff equal to 0.95 and the minimum standard deviation of variables set to 0.05. The final nonredundant set contains the following 24 descriptors: F_COUNT, P_COUNT, S_COUNT, CL_COUNT, BR_COUNT, ALOGP, MOLECULAR_POLARSURFACEAREA, NUM_H_ACCEPTORS, NUM_H_DONORS, NUM_ATOMS, NUM_HYDROGENS, NUM_POSITIVEATOMS, NUM_ROTATABLEBONDS, NUM_BRIDGEBONDS, NUM_RINGS, NUM_AROMATICRINGS, NUM_RINGASSEMBLIES, NUM_CHAINS, NUM_CHAINASSEMBLIES, NUM_STEREOBONDS, NUM_UNKNOWNSTEREOBONDS, NUM_ATOMCLASSES, LOGD, and MOLECULAR_WEIGHT.

3. Generation of r-SIFts. r-SIFt is a variation of structural interaction fingerprint (SIFt).[11,12] It incorporates the binding information about different variable R groups of a compound into the fingerprint. r-SIFt was specifically designed for processing and analyzing virtual screening results of combinatorial libraries. Both original SIFts and r-SIFts are binary bit-strings, representing the target−ligand interaction features of the binding site residues. The main difference between the original SIFt and r-SIFt lies in the meanings of the interaction bits that comprise the entire fingerprints. In the original SIFt, the bits represent the presence or absence of different types of interactions (contact, polar, hydrogen bonds, hydrophobic, etc.) occurring at each selected residue. In r-SIFt, the bits represent whether a certain R group or core fragment of the compound satisfies a contact interaction (i.e., within a distance threshold) with a particular protein residue. For example, suppose a ligand is comprised of core, R1, R2, and R3. Then for each binding site residue, we can use a four-bit-long binary string to represent the interaction pattern at this residue. These four bits represent whether the core, R1, R2, and R3 interact with this residue, respectively. If the core fragment is interacting with the residue, the "core" bit is turned on (1), otherwise it remains off (0). The same is true for the other bits. The final r-SIFt is constructed by

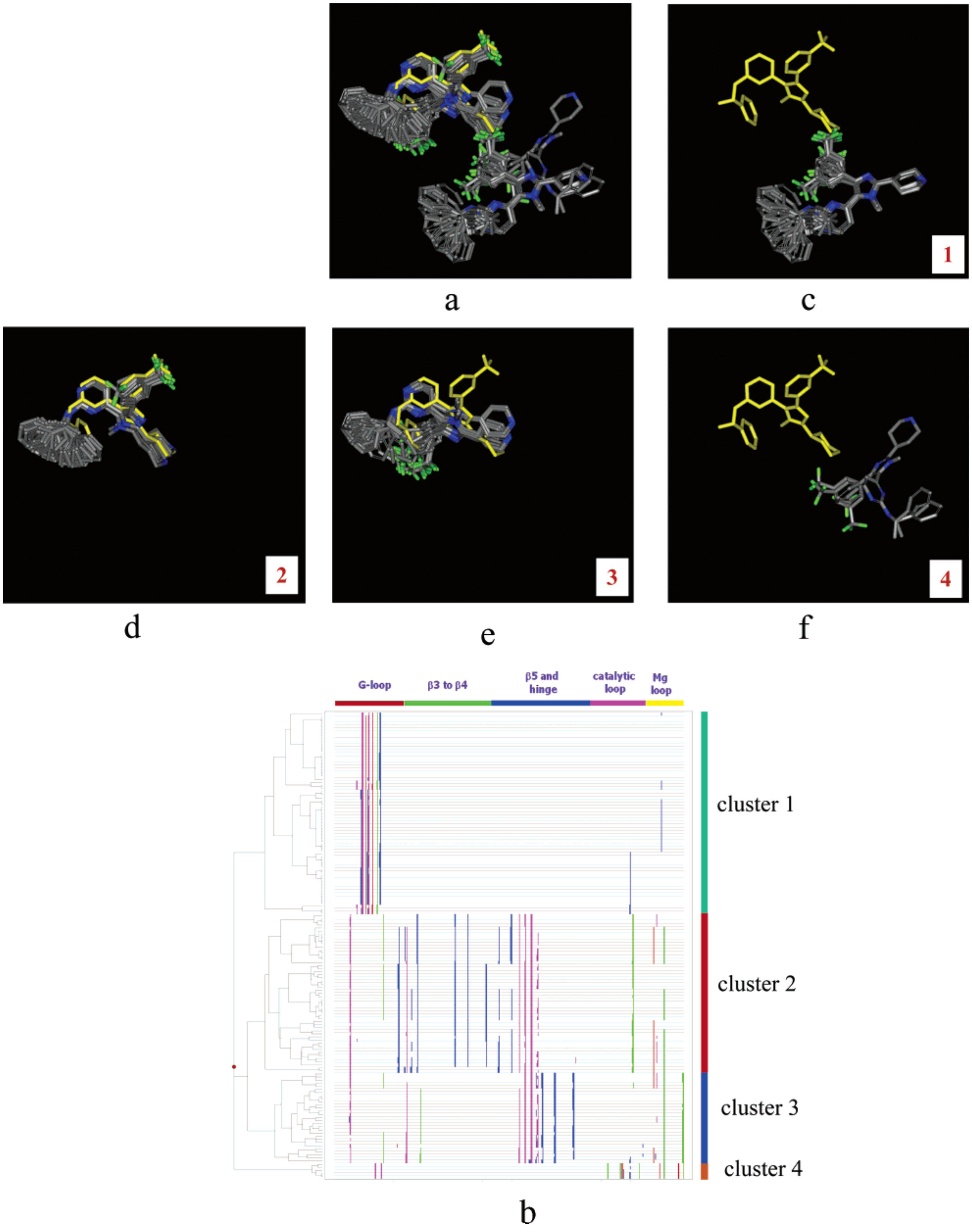

Figure 2. (a) Overlay of 150 poses of **1** docked onto the human p38 structure (PDB ID 1ouk). For comparison, the cocrystal structure of the molecule is shown in yellow while the docking poses are colored according to atom types. (b) Hierarchical clustering of the r-SIFts of 150 docking poses of **1**. Each r-SIFt is represented as one horizontal row in the heat map, and only on-bits (1) are shown. The interaction bits are colored accordingly to the respective molecular fragments (red, core; blue, R1; purple, R2; green, R3; see Figure 3 for R group definitions). The left side of the heat map shows the dendrogram of the hierarchical clustering result. r-SIFts in the heat map are rearranged according to the order given by clustering. Four major clusters (labeled 1−4) identified from the dendrogram are labeled on the right side of the r-SIFt heat map. The line of blocks above the heat map indicates the locations of the corresponding binding site residues in the protein. The residues are grouped into six different regions, as described previously.[11] For reason of clarity, the 56 residue numbers as well as the interaction bits are not displayed in the figure. The figure was generated using Spotfire.[33] (c−f) Overlay of the docking poses of each cluster (1−4), shown in the same reference frame as Figure 2a. The cocrystal structure of **1** is again displayed as a yellow stick model in each figure.

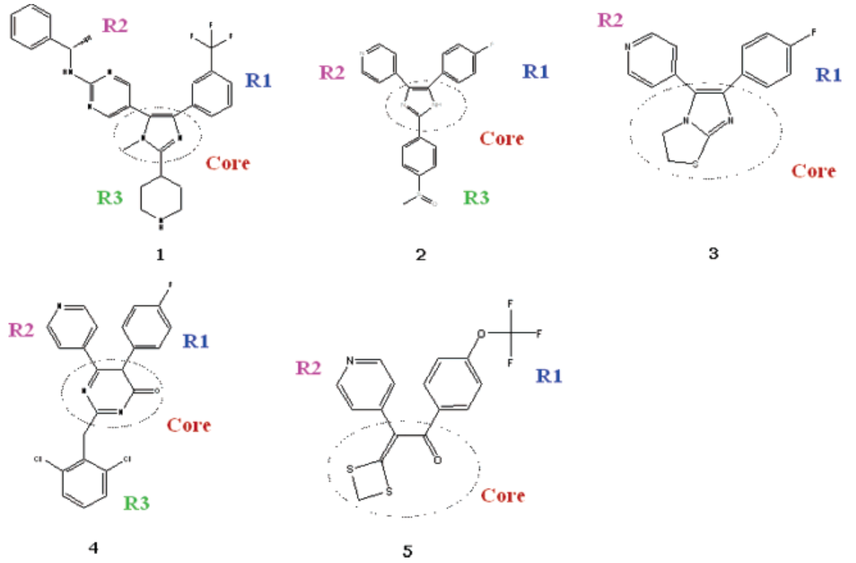

Figure 3. 2D chemical structures and R group definitions of compounds used in the study.

Table 1. Numbers of Native and Non-Native Compounds in Each Single R Group Variation Library

library	total	native	non-native
1-R1	2208	1428	780
3-R1	2442	266	2176
4-R1	1745	478	1267
1-R2	2450	352	1917
1-R3	2000	181	1819

sequentially concatenating the four-bit-long binary strings for all the binding site residues.

A panel of 56 residues of p38 previously identified as the kinase ligand-binding site was used as the reference frame for r-SIFt construction.[11] These residues are located in the vicinity of the ATP binding pocket in the cleft of the N-terminal and C-terminal domains, as well as at the substrate-binding site.

The implementation of r-SIFt used in this paper is based on the contact distance between the heavy atoms of a residue and different fragments of the ligands. There are different possible embodiments of r-SIFt. Here we used a four-bit-long binary bit string (and in the case of **3** and **5**, three bits, as they do not have R3) to represent the interactions involved in each binding site residue. Each bit represents whether a particular fragment (core, R1, R2, or R3) is within a certain distance cutoff (set to 3.5 Å) to the particular residue. If any heavy atom of a particular fragment is within 3.5 Å from any heavy atom of the residue, then this particular bit is turned on (1), otherwise this bit remains off (0). The final fingerprints were constructed by concatenating all these 56 small bit-strings together in ascending residue number order. The total length for each r-SIFt pattern is 56 × 4 = 224 bits, except for compounds in library 3-R1, in which R3 was absent. The length for r-SIFts in 3-R1 is 56 × 3 = 178 bits.

4. Analysis and Clustering of r-SIFts. We used the Tanimoto coefficient[30] as the similarity measurement between two r-SIFts. The Tanimoto coefficient (Tc) between two bit strings A and B is defined as

$$\mathrm{Tc(A,B)} = \frac{|\mathrm{A} \cap \mathrm{B}|}{|\mathrm{A} \cup \mathrm{B}|}$$

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.

For the 150 docking poses ensemble of **1**, as well as libraries 1-R1, 1-R2, and 1-R3, the cocrystal structure of the inhibitor **1** was used as the reference structure. As for **3** and **4**, the cocrystal structures were not available. We manually examined the top docking poses (with top FlexX scores) and selected a best pose for each inhibitor. These two best poses were consistent with the expected native binding modes as observed in the cocrystal structures of similar inhibitors (1ouk, 1a9u, 1bl6, 1bl7, 1bmk, 1ove, etc.) and satisfied all the conserved interactions with the target that were observed in other p38 structures.[12] We assumed them to be the correct binding mode and used these poses as the reference structures. We applied an agglomerative hierarchical clustering[31] to analyze and reorganize each library of poses, using Tanimoto coefficients as the similarity measurement. Clusters of protein−ligand complex structures were selected on the basis of the dendrogram of the r-SIFts.

In previous publications, we have shown that, by combining the SIFt-based approach and conventional scoring functions, one could achieve much better confidence in reproducing the true binding modes of the compounds and better library enrichment performance.[11,12] Our experience with docking known p38 inhibitors suggested that in many cases the best pose given by a conventional scoring function did not always adopt the native binding mode. However, a good placement with correct binding mode usually can be found among the top 10 poses. We have previously shown that for p38 inhibitors, retaining top 10 poses then selecting the poses with the best binding modes based on their SIFt similarities gave much better enrichment performance than using the conventional scoring function alone.[12] Here, we applied a similar strategy to process the docking results of the combinatorial libraries. We harvested the best 10 poses (with best Cscores) of each compound and then generated the r-SIFt patterns. Tanimoto coefficients were calculated against the r-SIFt of the respective reference structures (either the cocrystal structure or the best predicted pose as described above). The pose with the highest Tanimoto coefficient was selected as the best pose for this compound and used in subsequent ranking or hierarchical clustering.[31] All hierarchical clustering calculations of the r-SIFts were carried out using Spotfire.[33]

5. Construction of Decision Tree Classification Models. Hierarchical clustering grouped poses into different clusters according to their binding modes. By visual inspection, we could then easily identify the cluster in which compounds adopt the native binding mode. These compounds were classified as native, that is, they are “dockable”, because they were predicted by the docking program to be able to interact with the target molecule in a way similar to known active inhibitor(s). All other molecules, whose predictive binding modes are different from the native structure, were classified as in the non-native class.

After classifying the compounds, we generated decision tree models using CART (version 5, Salford Systems).[32] The nonredundant set of 2D descriptors was used as predictive variables and the binding mode class (native or non-native) as the target variable. The decision trees are comprised of a set of nodes and leaves (end nodes). Each node contains a bifurcation of path based on the value of a particular descriptor. We used 10-fold cross-validation, randomly assigning 90% of the data points as the training set and 10% as testing set. Equal weights were applied to both native and non-native classes. The performance of a model was measured by predictive accuracies for both classes in the training set and the test set.

Results

1. Organization of Docking Poses. We generated 150 poses by docking **1** onto p38 for r-SIFt analysis. Figure 2a shows the placements of these poses, which vary considerably in their binding modes. Hierarchical clustering of the r-SIFt patterns is shown in Figure 2b. The dendrogram clearly reveals four major clusters (clusters 1−4), each of which represents a distinct binding pattern (Figure 2c−f). This result demonstrates that r-SIFt is a very convenient and effective method for separating different binding modes.

In addition to its sensitivity to binding mode variations, r-SIFt renders a way for easy visualization and interpretation of how molecules are placed at the active site of the target molecule. Figure 2b displays the reorganized r-SIFt patterns as a heat map. Different types of interaction bits pertinent to different fragments of the compounds (core, R1, R2, and R3) are colored differently in the heat map. Since the bits in fingerprints were arranged in the same ascending residues number order, from this r-SIFt heat map one can easily reconstruct the overall orientation and position of the molecule at the active site, that is, which fragment of the molecule interacts with which region of the target molecule. Cluster 2 is the native cluster (Figure 2d). Within this cluster, the R1 groups (blue bits in the heat map) occupy and interact extensively with the hydrophobic pocket of p38, which is located at the back of the ATP binding site and is comprised of some residues in a sequence region spanning from $\beta 3$ to $\beta 5$ (including αC helix).[27] This binding information was revealed in the fingerprint heat map as blue bits (representing the R1 fragment) showing up in the region pertinent to the hydrophobic pocket. The R2 groups, on the other hand, interact with the adenine binding site in the hinge region; therefore, the majority of the purple bits (representing R2) show up in the hinge region. The R3 group (green bits) in this cluster touches the catalytic loop and the Mg-loop regions. Similarly, using this heat map as a guide, one can reconstruct the binding modes of other clusters and easily appreciate the differences among various groups, even without looking at their structures.

2. Comparison of r-SIFts of Different p38 Inhibitors. Furthermore, we carried out docking experiments using four known p38 inhibitors (**1**−**4**) and a compound (**5**) with no p38 inhibition activity. These compounds exhibit different chemical scaffolds (Figure 3). r-SIFt patterns were calculated for all docking poses and for each compound we selected three poses that displayed the best possible similarity scores against either the cocrystal structure or the respective best pose (i.e., the native binding mode). For **5**, we selected three poses with the highest Tanimoto coefficients against the cocrystal structure of **1** (1ouk), as it was difficult to predict the true binding mode of this noninhibitor. Hierarchical clustering results of these r-SIFt patterns are shown as a heat map in Figure 4a. The r-SIFt generated from the cocrystal structure of **1** is also displayed for comparison. Figure 4b−g shows the 3D structures of the poses of each compound within the same structural reference frame.

As shown in Figure 4a, not surprisingly, the r-SIFt patterns are first clustered together by each compound. Furthermore, the distance between two clusters in the dendrogram reflects the degree of similarity in the binding mode. In all four p38 inhibitors (**1**−**4**), the overall positions of the molecular fragments within their r-SIFts are consistent. In most of the cases, the R2 group (purple bits) is in contact with the hinge region, whereas the R1 group (blue bits) is highly concentrated in the hydrophobic pocket region. This result shows that different p38 inhibitors bind to the target molecule with a very consistent overall interaction pattern. **5**, on the other hand, displays a completely different binding mode and is the most distant from other inhibitors in the dendrogram.

A more detailed investigation of the r-SIFt patterns reveals some degree of variation between different known inhibitors. For example, the R2 group of **1** (purple bits in Figure 4a) shows more extensive interactions in the second half of the hinge region (around residue 110) than other inhibitors. Such extensive contact between **1** and the hinge has been previously observed and rationalized.[12,15] In addition, **1** exhibits more interaction points than other compounds in the hydrophobic region. This difference can be rationalized by the fact that it has a bulkier trifluorobenzene R1 group as opposed the smaller 3-fluorophenol R1 of the others. In addition, the interactions between the R2 of **4** and the hinge region are relatively sparser than for other molecules. As seen from the structures, the **4** poses predicted by our docking experiments move slightly away from the hinge, so that the carbonyl at the core can make hydrogen-bonding interaction with Lys-53. The relative distance between different compounds correlates well with chemical similarity, with **2** and **3** being very close to each other (their R1 and R2 are identical and cores are similar), while **1** and **5** (chemically more dissimilar) are farther apart in the dendrogram.

These p38 inhibitors adopt a generally similar binding pattern when binding to the same target molecule, and their binding modes are highly correlated to their chemical structures. The way they bind to a target molecule is dictated by their own chemical properties. Conversely, given a particular target, we can expect that only molecules exhibiting certain physical and chemical properties are able to bind to the target with desirable binding mode. Therefore, finding the rules for such chemical feature subspace would be highly valuable in rational library design.

3. Analysis of Combinatorial Libraries. To search for the rules governing the behaviors of the compounds within a target, we first enumerated five combinatorial libraries and used r-SIFt to help investigate their "dockability", that is, whether they were able to dock onto the target with an expected binding mode. After docking the compounds and generating r-SIFts, we carried out hierarchical clustering analyses to separate different binding modes. Figure 5a shows the organization of these r-SIFt patterns of library 1-R1. The first major cluster (illustrated in green) is the native cluster in which the relative positions and orientations of the molecules in the cluster are similar to those observed in the cocrystal structure (1ouk). Examples of the compounds in this native cluster are shown in Figure 5b. The rest of the library was labeled as non-native (shown in red). Figure 5c shows the chemical structures of some example molecules in both native and non-native clusters. We should note that all of these examples shown in Figure 5c had high docking scores; therefore, using conventional docking score alone we would not have been able to effectively separate the ones with native binding. A few molecular descriptors show modest correlation with the r-SIFt classification. For example, in the native cluster, the molecular surface areas of the R1 groups in general tend to be smaller than in the non-native cluster (data not shown). This difference can be rationalized as the size of the p38 hydrophobic pocket precluding a very large R1 group from occupying the pocket.

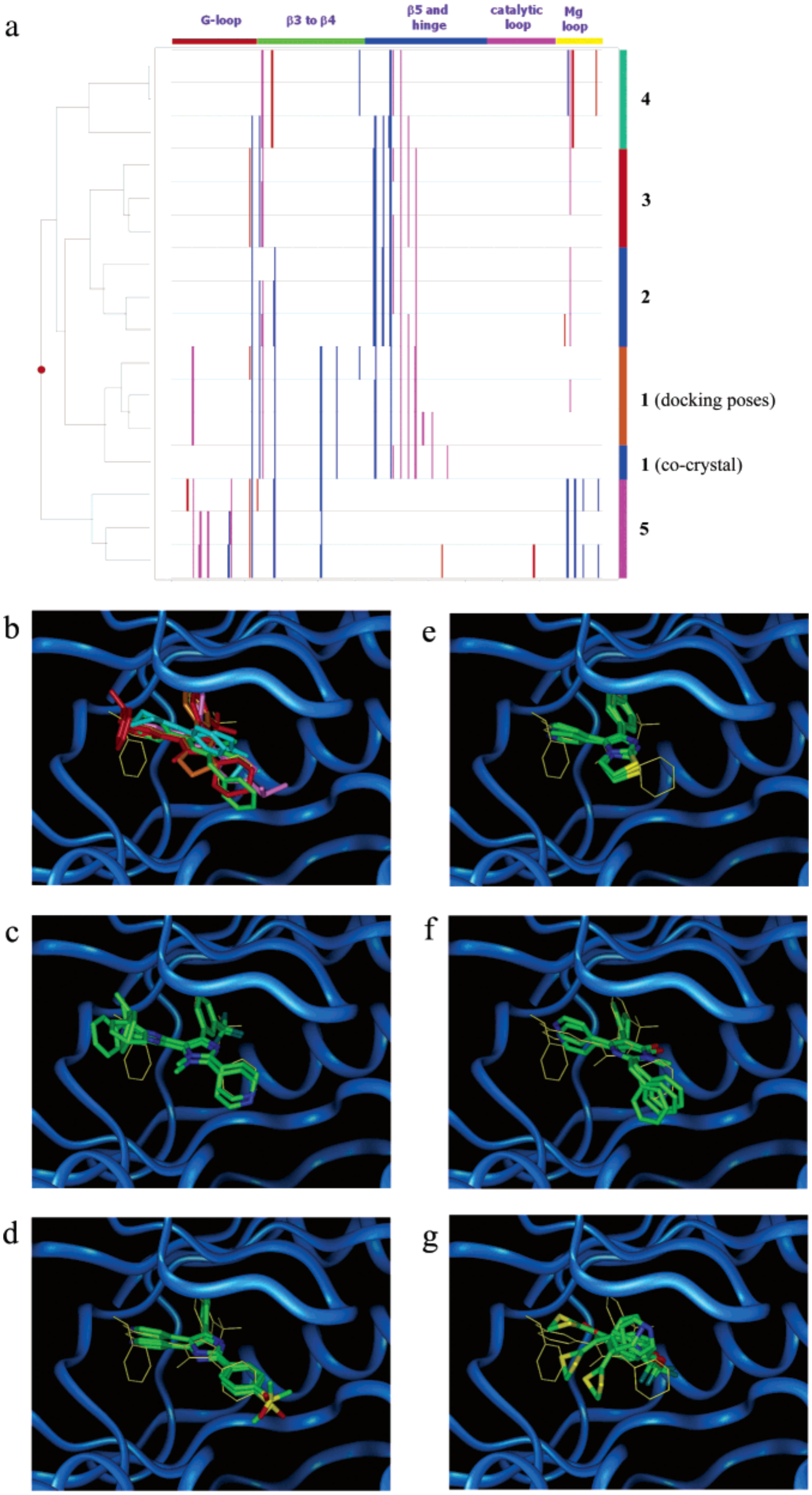

Figure 4. (a) Hierarchical clustering of r-SIFts derived from the docking poses of five different compounds docked into the p38 structure (1ouk). The bit-coloring scheme and structure layout are identical to those in Figure 2b. For each compound, three poses with the best r-SIFt Tanimoto coefficients were chosen and analyzed. Since **3** does not contain R3, all the r-SIFt patterns on display were constructed by omitting all the R3 bits (if present). For comparison purposes, the r-SIFt pattern of the cocrystal structure of **1** is also included. (b) An overlay of the best docking pose of each of the five molecules, within the same active site of the target molecule structure. The cocrystal structure of **1** is shown as a thin yellow line model for comparison. (c−g) Structures of the docking poses of each compounds (three poses per molecule) used in Figure 4a, shown in the same reference frame as in part b. The compounds are (c) **1**, (d) **2**, (e) **3**, (f) **4**, and (g) **5**.

However, neither the size nor the hydrophobicity of the R1 groups alone (or the combination of these two) was able to successfully explain the classification variance. Therefore, a predictive model that involves more complicated combination of different descriptors was required. The CART decision tree method was used to build such classification models.

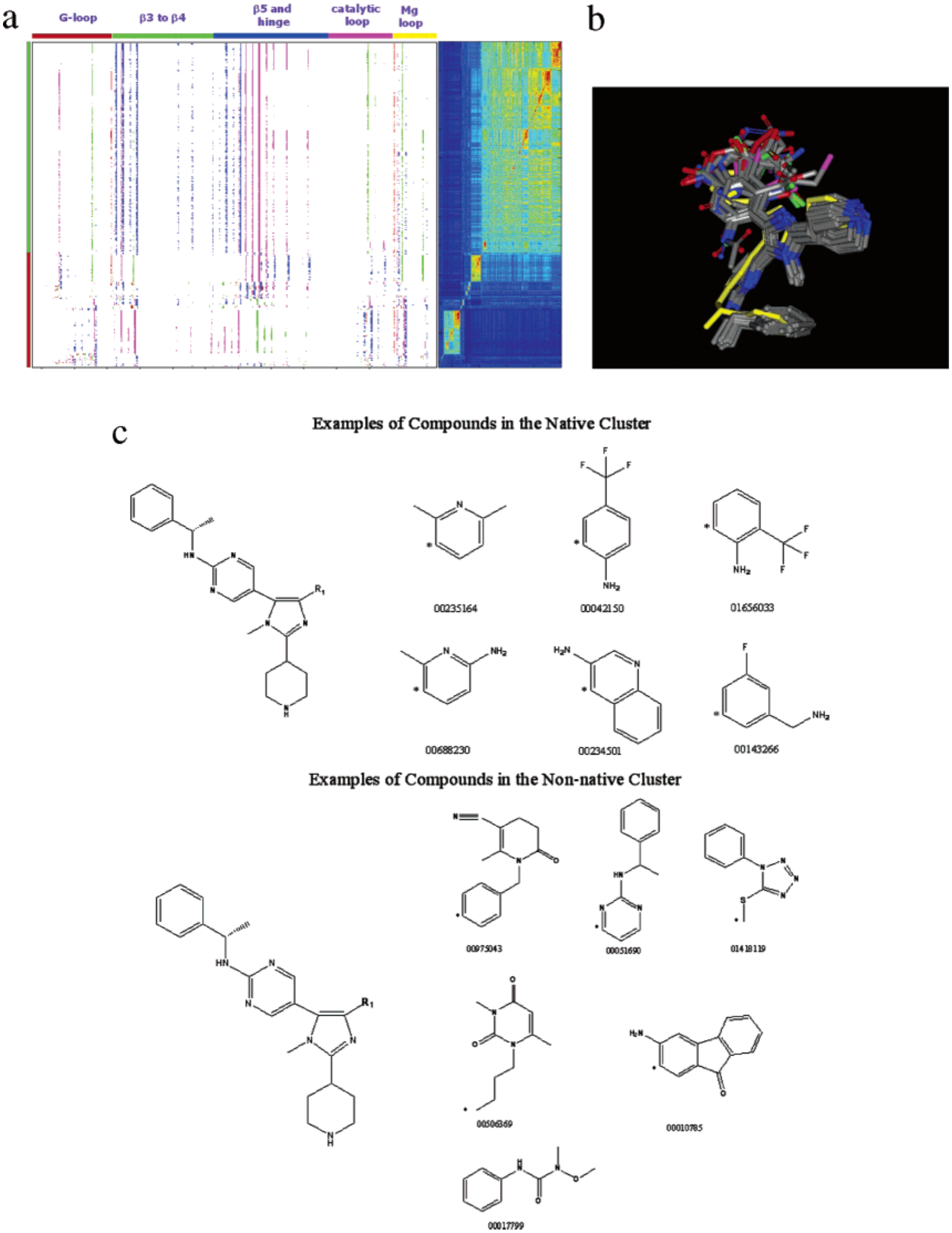

Figure 5. (a) Classification of the 1-R1 library compounds based on their r-SIFt similarities. The coloring scheme is the same as in Figure 2b. For clarity, we replaced the dendrogram as shown in Figures 2b and 4a with a Tanimoto coefficient distance matrix of the r-SIFt patterns. The compound order in the distance matrix matches that in the SIFt heat map, and the coloring gradient in the distance matrix corresponds to the values of the Tanimoto similarity score, from dark red (highest similarity) to dark blue (least similar). The compounds that display the native binding mode similar to the cocrystal structure (Figure 5b) in which the R1 groups are correctly located in the hydrophobic region are labeled as a “native cluster”, and the rest of the compounds are labeled as “non-native”. (b) The 3D structures of 200 example compounds in the native cluster. The cocrystal structure of **1** is shown as a yellow stick model. (c) Examples of compounds in native and non-native clusters. The R1 attachment points are labeled with an asterisk. The numbers correspond to the unique IDs in the original reagent library.

We generated a decision tree model for each of the five combinatorial libraries, using a nonredundant set of their 2D molecular descriptors as predictive variables. Figure 6 shows the optimal decision tree model for library 1-R1. The CART program also produced a sorted list of descriptors based on their levels of importance. Descriptors that are pertinent to the size, shape, polarity, and hydrophobicity of the R1 group, such as total number of atoms, total surface area, polar surface area, molecular weight, and log *D*, are among the most informative decision tree splitters. This finding is consistent with the fact that the size, shape, and hydrophobic nature of the hydrophobic pocket impose restrictions onto the R1 groups such that only those compounds with the right size, shape, and hydrophobicity were able to fit in the well-defined site with the desired binding mode.

The performances of these decision tree models were evaluated by the prediction accuracies for both native and non-native

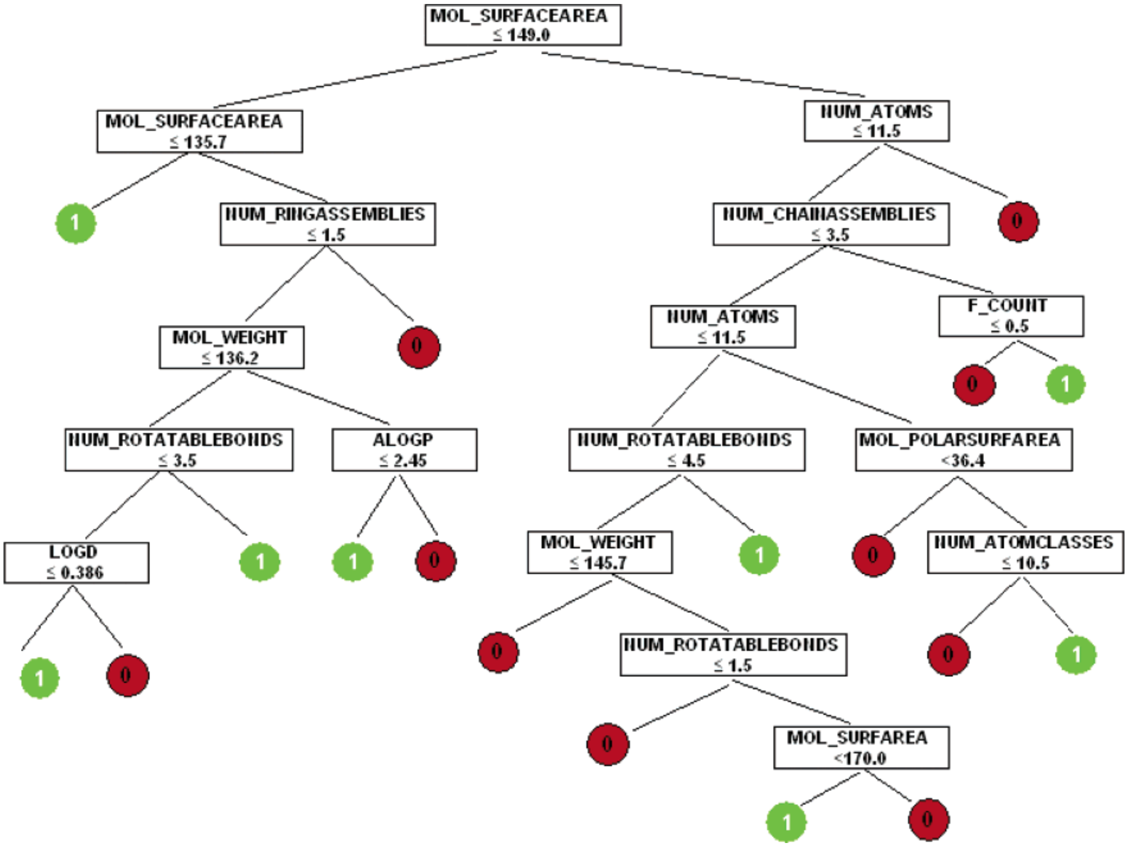

Figure 6. A decision tree predictive model for the 1-R1 library. The green end nodes (labelled "1") are predicted native cluster; whereas the red end nodes (labelled "0") are non-native cluster.

Table 2. Performances of the Decision Tree Predictive Models[a]

	training set		test set	
library	native	non-native	native	non-native
1-R1	80	84	80	77
3-R1	78	80	77	79
4-R1	83	74	73	70
1-R2	71	72	70	71
1-R3	64	75	57	68

[a] Numbers are the percentage of molecules in either native or non-native classes that were accurately predicted by the decision tree models. Ten percent of each original data set was randomly selected and set aside as the validation test set and was not used in model building.

Table 3. Cross-Library (R1 variation only) Model Prediction Accuracies[a]

	target library					
model	1-R1		3-R1		4-R1	
library	native	non-native	native	non-native	native	non-native
1-R1	–	–	78	71	90	50
3-R1	74	74	–	–	83	56
4-R1	81	61	82	62	–	–

[a] Each of the predictive models generated from three different R1 libraries (i.e., model libraries) was used to test against the other two R1 libraries (i.e., target libraries). The numbers represent the percentage of molecules in the target libraries that were correctly predicted by the model.

classes. The results are summarized in Table 2. We used 10-fold cross-validation during the construction process, using 90% of the data points (randomly selected) each time to build the model while setting aside 10% of the data as test set for validation. The accuracies against the test data sets left aside during the decision tree construction is a better performance indicator.[32] Most of these models gave reasonably good and balanced performances, with accuracies (against test sets) in the range of 70–80% for both native and non-native classes of molecules.

The three R1 libraries were derived from different scaffolds. Since the variable R1 groups in these libraries all target the same hydrophobic binding pocket, it is reasonable to expect that the rules derived from these libraries are closely related to each other. To test this hypothesis, we applied each decision tree model to predict the other two R1 libraries. The cross-library prediction results are summarized in Table 3. 1-R1 and 3-R1 are interchangeable, with their cross-library prediction accuracies remaining 71–78% for both classes of molecules, a performance comparable to their self-prediction accuracies (Table 2). Interestingly, all cross-library prediction performances containing the 4-R1 library show different accuracies for native and non-native classes: the native classes can be predicted more accurately (81–90%) than the non-native molecules (only 50–62%).

To test the expandability of these predictive models, we further regenerated decision trees by randomly setting aside 25% (500 compounds) of the original library as the evaluation set. Models were built using the remaining 75% of the data, with exactly the same parameter settings and the same 10-fold cross-validation. We then applied each model to test the respective evaluation set that was never used in the model building process. The prediction accuracies were all comparable to those shown in Table 2 (data not shown), indicating that the models are expandable and therefore can be used to filter large libraries.

Furthermore, we applied the same approach to investigate larger libraries of true combinatorial nature, i.e., varying more than one position simultaneously. To determine the suitable size for a test library, we first investigated the performance of an R group decision tree as a function of the sample size. We

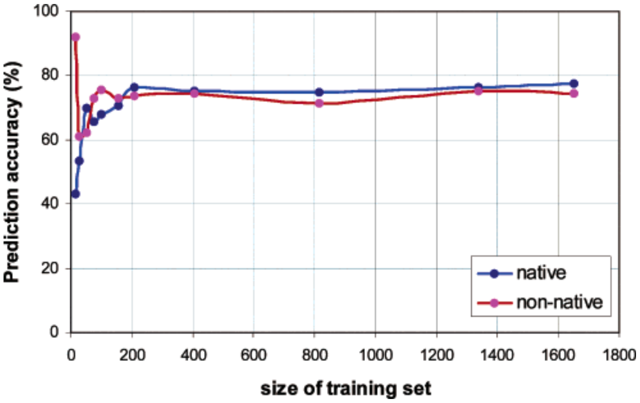

Figure 7. Effect of training set size on the performance of predictive models of the 1-R1 library. Decision tree models were generated using training sets of various sizes, and the performances were cross-validated against an evaluation set of 500 compounds randomly selected and set aside during model construction. The models are unstable when the training set sizes are below 100, and they reach stable plateaus after 200.

randomly set aside 25% of the 1-R1 library as an evaluation set. Several decision tree models were generated, each using different numbers of compounds randomly selected from the remaining 75% of the library. The performances of these models were measured against the same evaluation set, and the results are plotted in Figure 7.

Interestingly, the performance of the model (for both native and non-native classes) starts to level off after a sample size of ~100. Increasing the training set size beyond 200 actually leads to little gain in predictive accuracy. This result implies that, for the purpose of the r-SIFt analysis, if selected randomly, a training set of modest size is able to effectively cover most of the chemical subspace of the whole library.

Using **1** as template, we enumerated a combinatorial library of 10 000 compounds with 100 R1 and 100 R2 groups, while keeping the core and R3 fragment unchanged. We wanted to construct decision tree predictive models for R1 and R2 from this full matrix combinatorial library in which more than one R group was varied simultaneously, to see how different variable groups affected each other. Both R1 and R2 used the same monomer library of 100 commercially available aryl bromides, which had not been used in the construction of the five independent libraries described earlier. A sample size of 100 has been shown to be able to generate a reasonably stable predictive model (Figure 7).

This R1(100) × R2(100) training library was then docked onto the 1ouk structure and the same r-SIFt clustering analysis was carried out. From the native cluster, we gathered a list of dockable R1 and R2 groups. An R1 group was considered dockable if at least one compound containing this group as R1 adopted the native binding mode. The same was true for R2. The rest of the R1 or R2 groups were classified as undockable, i.e., no compound with this R group was found in the native cluster.

On the basis of these classifications, two separate decision tree models were built for R1 and R2 groups, respectively, using the same parameter settings described before. The performances of these two models were measured by the predictive accuracies against their respective cross-validation testing sets. For the R1 model, the accuracies were 67% for the native class and 72% for the non-native class; for the R2 model, accuracies were 66% for the native class and 75% for the non-native class. These predictive accuracies were all consistent with those generated from independent libraries with single variable point (Table 2). In fact, the independent R1 and R2 models were both comparable to the new models: the old R1 and R2 models were able to predict the dockability of the R groups in the 100 × 100 library with 88% and 92% accuracies for the native class, and 59% and 61% for the non-native class, respectively.

The results from the 100 × 100 combinatorial library suggested that different variable groups can be treated independently. To further test the hypothesis, we enumerated a library of 1 million compounds using **1** as template, varying R1 (100), R2 (100), and R3 (100) simultaneously. A subset of 3580 compounds was randomly sampled from this library and docked onto the 1ouk structure. We then used the R1, R2, and R3 models that were built from the respective single-point variation libraries to classify each of the variable groups in the subset. Compounds whose R1, R2, and R3 were all classified as good groups by their respective models were considered dockable. A total of 88 compounds out of 3580 passed all three filters. Docking and r-SIFt analysis showed that 29 (0.8% of 3580) compounds were able to adopt the native binding mode, among which 17 were correctly predicted by the combination of R1, R2, and R3 models. Therefore, 58.6% (17/29) of the true dockable compounds were correctly predicted; for the non-native class the success rate was 3480/3551 = 98%. We can also use the enrichment factor (EF),[35] which represents the increased concentration of dockable compounds in the selection pool, as another measurement for the success of library focusing. The EF compared to the original library is (17/88)/(29/3580) = 23.8. Hence, using our r-SIFt based approach, we were able to achieve a ~24-fold enrichment in the concentration of dockable molecules.

Discussion and Conclusion

Reagent selection for automated parallel synthesis (APS) and, in general, combinatorial library design is a task encountered routinely during the course of lead discovery and optimization. Typically, a subset of reagents appropriate for the underlying chemical reaction must be selected from an often vast and diverse list of commercially available reagents. Ideally, reagent selection can be informed by available structural information to generate target-focused libraries. Although it is possible to enumerate and dock the library in order to arrive at a subset of optimal reagents, this approach quickly runs into a limitation due to the size of the combinatorial libraries.

In this paper, we present a general workflow for designing target-focused chemical libraries, where information on the desired binding mode can be directly embedded into the reagent selection process. Key to this approach is to use the r-SIFt method to effectively classify compounds on the basis of whether they can interact with the target while satisfying desired binding patterns and then use machine learning techniques to build filtering rules that can be applied to large libraries.

The overall flowchart of this strategy is illustrated in Figure 8. This method takes advantage of the ability of SIFt (including r-SIFt) to quickly analyze and organize large amounts of structural data and to efficiently identify compounds consistent with known binding modes from large docking data sets. The strategy involves the following steps: (1) select a small pilot library from the original large combinatorial library, with maximized diversity; (2) calculate 2D descriptors of the whole library of compounds; (3) dock this small library onto the target molecule structure; (4) calculate r-SIFt or traditional SIFt patterns for the docking poses; (5) analyze and cluster the poses on the basis of their r-SIFt patterns; (6) on the basis of the SIFt

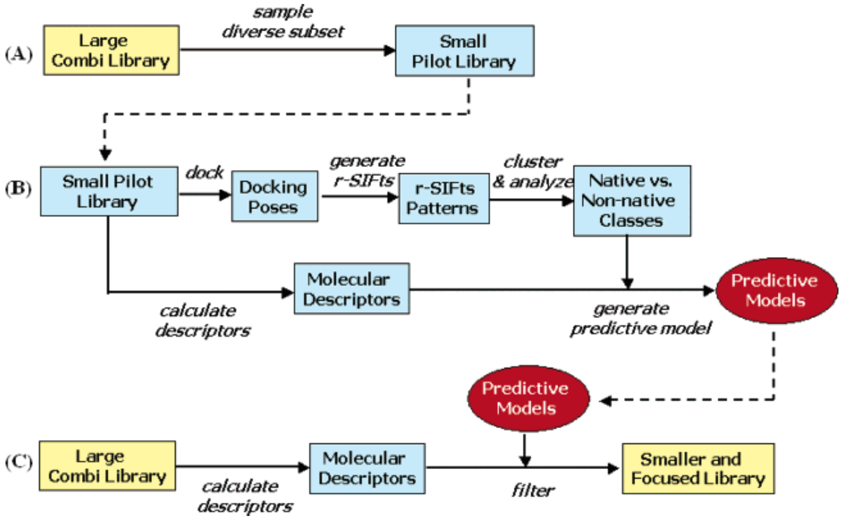

Figure 8. An r-SIFt-based chemical library focusing workflow.

analysis results, classify compounds into native and non-native groups, according to whether they are able to bind to the target molecule with the desired binding mode or satisfy some predefined interactions; (7) build predictive models based on the above classifications, using the 2D descriptors of the R groups as predictive variables; and (8) apply this predictive model to filter the original large combinatorial library.

r-SIFt is a new variation of our SIFt specifically designed for combinatorial libraries that offers several advantages for library design. When represented as a heat map, the r-SIFt patterns provide a convenient tool for direct visualization of how the variable R groups defining the library interact with the target molecule. More importantly, we have demonstrated that the resultant predictive models for unique R groups are separable and can be applied independently of each other to select reagents for synthesis. The independence of the models reduces library focusing to filtering at the reagent level in a series of selection steps for each R group, thereby avoiding the combinatorial explosion faced when focusing the enumerated compound library. Hence, much larger libraries can be focused than would be impossible via enumeration and subsequent structure- or ligand-based focusing.

To investigate the performance of r-SIFt, we applied the method to the problem of focusing a large APS combinatorial R1 × R2 × R3 library down to an optimal set of commercially available reagents. The goal of this example was to identify a subset of reagents for R1, R2, and R3 that would be enriched for compounds that bind in the desired binding mode to p38. On the basis of r-SIFt classification, models for reagents were generated from libraries where each R group was varied independently, while the others were kept fixed. The independent models were found to have good accuracy at predicting the reagents that would bind as desired. To test the validity of treating R1, R2, and R3 independently, we carried out the r-SIFt procedure on a library where R1(100) × R2(100) were varied simultaneously. The accuracies of the R1 and R2 models derived from the coupled library were found to be comparable to those obtained from the independent R group libraries. Finally, to test the accuracy of our approach on the full combinatorial library, an R1(100) × R2(100) × R3(100) library was enumerated and a test subset focused using the independent R group models. Using r-SIFt focusing, the reagents selected were enriched by 24-fold for good binders.

To capture the receptor R group preferences, the predictive models are necessarily complex. However, some intuitive insights can be gleaned from the models. In the previous example, for instance, the R1 model demonstrates a preference for compact planar, aromatic/heteroaromotic groups with small substituents. These criteria are consistent with having to bind in the well-defined hydrophobic pocket of p38. In contrast, the R3 model (not shown), which corresponds to binding in the P-loop region, selects primarily for 1,4-substituted phenyl (or six-membered heteroaromatic) groups. Such a pattern makes sense at the R3 position, given the prevalence of similar moieties for p38 inhibitors.[24] Finally, the R2 model (not shown) for the adenine binding site substituent favors rings containing a hydrogen bond acceptor and selects for longer chains with more rotatable bonds than either R1 or R3. The R2 model trends are consistent with binding at the hinge region of p38 that is open and can tolerate extended chains leading to solvent. Finally, the selection for hydrogen bond acceptor containing groups is notable, because the critical hydrogen-bonding interaction with the hinge[11,12] was not specified as an r-SIFt constraint.

The task inherent in target-based focusing approaches is to define, search, and identify the small chemical subspace of compounds that can fit into the ligand-binding site with an expected binding mode. We use r-SIFt fingerprints to define the target constraints and apply decision tree models to efficiently search the chemical space defined by all possible R groups for good binders. All decision tree filters work well for their corresponding libraries. Moreover, the predictive models are based on 2D descriptors of the reagents and can therefore be used as molecular filters to sift through very large libraries, even when the cores and scaffolds may be different.

r-SIFt can be a valuable approach in applying binding mode constraints when the expected binding mode is known. The method is a variation and extension of the traditional SIFt method, incorporating the binding information of different variation points in a combinatorial library into the fingerprints. As a result, the 3D library can be visualized and analyzed in the target-binding site on the basis of how the variable R groups in each library member interact with the target. The usefulness of r-SIFt also lies in its flexibility. Depending on the library analysis requirements, different variants can be defined by selecting and incorporating various types of binding information into the fingerprints.

One should keep in mind that the version of r-SIFt described in this paper only provides information about the interaction characteristics of ligand variable R groups with a target-binding site. The patterns do not, however, contain more detailed information about what kinds of interactions (hydrophobic, polar, hydrogen bonds, etc.) are involved, as is provided by the traditional SIFt[11] and the SIFt profile (p-SIFt)[12] approaches. However, it is possible to apply more than one type of SIFt in the library design process. For example, we can use r-SIFt to identify a set of optimal R groups from a large pool of available reagents, enumerate the much smaller library, and then apply traditional SIFt to further search for molecules making specific interactions with particular residues/subregions. Such layering of interaction constraints would generate a pool of native molecules that have the potential to be more specific and selective.

In summary, the r-SIFt method described in this paper offers a sensitive and efficient technique to discriminate the binding characteristics of combinatorial libraries and can be used to define a target-based reagent selection strategy for library design. The library focusing workflow presented in this paper scales linearly with the number of R groups (i.e., complexity $O(N_1 + N_2 + N_3 + ... + N_M)$, where M is the number of variable groups, and N_i is the total number of reagents for the ith variable group), because the predictive models can be derived by treating the R groups as independent of each other. As a result, large vendor lists of reagents can be searched for each R group independently without the need to enumerate the entire library (i.e., complexity $O(N_1 \times N_2 \times N_3 \times ... \times N_M)$), vastly increasing the size of the chemical space that needs to be searched. In addition, the r-SIFt approach offers an efficient means to rank combinatorial libraries, for example, from two vendors, on the basis of the predicted enrichment generated from the R group models. In conclusion, we believe that r-SIFt is a unique and novel approach for visualizing and focusing combinatorial libraries on the basis of their predicted interaction patterns.

Acknowledgment. The authors would like to thank Anuj Patel and Aravind Prasad, for their help in test compound selection and the R group fragmentation protocol; Gaurav Narale, for his help in incorporating the r-SIFt analysis components in Pipeline Pilot; and Drs. Rainer Fuchs, Donovan Chin, Herman Van Vlijmen, Alexey Lugovskoy, Jennifer Campbell, and James Dowling, for their insightful discussions and critical reviews of the manuscript. Finally, we thank the reviewers of this paper for their insightful comments and suggestions.

References

(1) Ghose, A. K.; Viswanadham, V. N. Combinatorial library design and evaluation: Principles, software tools and applications in drug discovery. Marcel Dekker: New York, 2001.

(2) Valler, M. J.; Green, D. Diversity screening versus focused screening in drug discovery. *Drug Discov. Today* **2000**, *5*, 286−293.

(3) Jamois, E. A.; Lin, C. T.; Waldman, M. Design of focused and restrained subsets from extremely large virtual libraries. *J. Mol. Graph. Mod.* **2003**, *22*, 141−149.

(4) Bravi, G.; Green, D. V. S.; Hann, M. M.; Leach, A. R. PLUMS: A program for rapid optimization of focused libaries. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1441−1448.

(5) Teckentrup, A.; Briem, H.; Gasteiger, J. Mining high-throughput screening data of combinatorial libraries: Development of a filter to distinguish hits from nonhits. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 626−634.

(6) Stanton, R. V.; Mount, J.; Miller, J. L. Combinatorial library design: Maximizing model-fitting compounds within matrix synthesis constraints. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 701−705.

(7) Wright, T.; Gillet, V. J.; Green, D. V. S.; Pickett, S. D. Optimizing the size and configuration of combinatorial libraries. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 381−390.

(8) 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.

(9) Blundell, T. L.; Jhoti, H.; Abell, C. High-throughput crystallography for lead discovery in drug design. *Nat. Rev. Drug Discovery* **2002**, *1*, 45−54.

(10) McGregor, M. J.; Muskal, S. M. Pharmacophore fingerprint. 2. Application to primary library design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117−125.

(11) 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.

(12) Chuaqui, C.; Deng, Z.; Singh, J. p-SIFt: Interaction profiles of protein kinase−inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121−133.

(13) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual screening−An overview. *Drug Discovery Today* **1998**, *3*, 160−178.

(14) Xu, H.; Agrafiotis, D. K. Retrospect and prospect of virtual screening in drug discovery. *Curr. Top. Med. Chem.* **2002**, *2*, 1305−1320.

(15) 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 p38a MAP kinase quinazoline and pyridol-pyrimidine inhibitor specificity. *Nat. Struct. Biol.* **2003**, *10*, 764−769.

(16) 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.

(17) SYBYL (version 6.9), Tripos, Inc., St. Lois, Missouri, 63144.

(18) 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.

(19) 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.

(20) Gohlke, H.; Hendlich, M.; Klebe, G. Knowledge-based scoring function to predict protein−ligand interactions. *J. Mol. Biol.* **2000**, *295*, 337−356.

(21) 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.

(22) 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*, 5100−5109.

(23) 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*, 1117−1128.

(24) 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.

(25) WIPO patent WO 0031063A1.

(26) ACD: Available Chemical Directory (version 2004.2), MDL Information Systems: San Leandro, CA.

(27) Radzio-Andzelm, E.; Taylor, S. S. Three protein kinase structures define a common motif. *Structure* **1994**, *2*, 345−355.

(28) Pipeline Pilot (version 3.0), Scitegic Inc., San Diego, CA.

(29) Whitley, D. C.; Ford, M. G.; Livingstone, D. J. Unsupervised forward selection: A method for eliminating redundant variables. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1160−1168.

(30) Willet, P. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.

(31) Dubes, R.; Jain, A. K. Clustering methodologies in exploratory data analysis. *Adv. Comput.* **1980**, *19*, 113−228.

(32) Steinberg, D.; Colla, P. CART: Tree-structured nonparametric data analysis. Salford Systems, San Diego, CA, 1995.

(33) Spotfire Decisionsite (version 7.3), Spotfire, Inc., Somerville, MA.

(34) Hanks, S. K.; Hunter, T. The eukaryotic protein kinase superfamily: Kinase (catalytic) domain structure and classification. *FASEB J.* **1995**, *9*, 576−596.

(35) Pearlman, D. A.; Charifson, P. S. Improved scoring of ligand−protein interactions using OWFEG free energy grids. *J. Med. Chem.* **2001**, *44*, 502−511.

JM050381X