Position Specific Interaction Dependent Scoring Technique for Virtual Screening Based on Weighted Protein−Ligand Interaction Fingerprint Profiles

Ravi K. Nandigam,*,[†] Sangtae Kim,[†] Juswinder Singh,[‡] and Claudio Chuaqui[‡]

School of Chemical Engineering, Purdue University, West Lafayette, Indiana, and
Biogen Idec, Cambridge, Massachusetts

Received December 22, 2008

The desire to exploit structural information to aid structure based design and virtual screening led to the development of the interaction fingerprint for analyzing, mining, and filtering the binding patterns underlying the complex 3D data. In this paper we introduce a new approach, weighted SIFt (or w-SIFt), extending the concept of SIFt to capture the relative importance of different binding interactions. The methodology presented here for determining the weights in w-SIFt involves utilizing a dimensionality reduction technique for eliminating linear redundancies in the data followed by a stochastic optimization. We find that the relative weights of the fingerprint bits provide insight into what interactions are critical in determining inhibitor potency. Moreover, the weighted interaction fingerprint can serve as an interpretable position dependent scoring function for ligand protein interactions.

INTRODUCTION

A key to achieving success in rational drug design is the ability to exploit available structural information to effectively feed into design-make-test cycles. For target classes such as protein kinases, an overwhelming number of 3D X-ray and NMR structures can make it difficult to systematically extract useful information that can be applied to inhibitor design or to inform virtual screening. To this end, a variety of approaches have been developed to capture structural knowledge (beyond simple visual inspection) ranging from 2D ligand filters and pharmacophores through to 3D pharmacophoric features and interaction constraints that can be used to limit poses generated from ligand receptor docking.

Structural Interaction Fingerprint (SIFt) is a method developed by Deng et al.[1] for efficiently representing and analyzing the binding interactions in protein−ligand complexes. SIFt first involves identifying key residues in the binding pocket of the complex and then examining the interactions made by the ligand at each key residue, followed by representing the information of the interactions as a unique fingerprint. Each residue is assigned a fixed set of bits in the SIFt that characterize the interaction between the ligand and the residue The application of SIFt to visualize, organize, analyze, and simultaneously mine databases of protein−ligand complex structures has been elaborately discussed in refs 1 and 2. The conservation of interactions for a set of structures can be represented by an interaction profile where each element measures the average occupancy (or frequency) of a bit over the set of fingerprints.[2] In contrast, the weighted interaction profiles introduced in this paper incorporate an empirically determined weight fit from inhibitor potency data. The profile weights are determined such that the fingerprint similarity between docked poses and the weighted profile is in effect a residue-specific QSAR based on the relative importance of ligand−receptor interactions for determining potency.

Interaction fingerprints and profile-based methods have been applied to virtual screening, library design, and the analysis of large numbers of X-ray structures to identify interaction patterns that may influence inhibitor potency and selectivity. Moreover, because binding information is encoded in a 1D fingerprint, advanced filtering, clustering, and machine learning methods may be applied to identify patterns underlying the binding data, thereby enhancing the ability to make useful implications that are not apparent by looking at individual structures. The evolution of interaction fingerprint and profile approaches and their application to docking, scoring, and the analysis of ligand−receptor interactions has been comprehensively reviewed recently by Brewerton[3] and will not be further reviewed here.

Weighted Interactions Profiles. SIFt does not encode all binding information between the protein and the ligand in a complex, because all interactions are treated simplistically and identically. Characterizing ligand receptor interactions by residue specific on/off bits is obviously a gross simplification of molecular recognition that while very useful for filtering, clustering, and mining ligand poses does not convey the full information about the binding interactions at the residue. As an example, it is well-known that in kinases interactions with the hinge region are critical for binding compared to residues at other regions. Interaction fingerprints and profiles do not treat the information pertaining to the hinge region residues differently from that of any other residue. Likewise, the strength of interactions at a given residue is not captured in the unweighted interaction profiles.

* Corresponding author phone: (972)603-8796; e-mail ravi.nandigam@aspentech.com.
[†] Purdue University.
[‡] Biogen Idec.

10.1021/ci800466n CCC: $40.75 © 2009 American Chemical Society
Published on Web 05/05/2009

SIFt 1	1	0	0	1	1	0	0	1	0	1
SIFt 2	1	1	0	0	1	0	1	0	0	1
SIFt 3	1	1	0	0	1	0	1	0	0	1
SIFt 4	1	1	0	0	1	0	1	1	0	1
SIFt 5	0	0	0	1	1	0	0	1	0	1
W-Profile	w_1	w_2	0	w_4	w_5	0	w_7	w_8	0	w_{10}

Figure 1. Illustration of weighted profile for a set of interaction fingerprints. The profiles reported in ref 2 are not weighted and are the simple average of the bits at each position in the profile.

In order to incorporate a more robust representation of the ligand receptor interaction into the interaction profile, we propose to introduce a weight to each interaction bit in profile that captures the relative importance of that particular bit for binding. This extension of SIFt, called weighted SIFt or w-SIFt, in addition to efficiently accounting for interactions of varying degree of importance, expands the possible range of applications to which SIFt can be useful. In this work, we demonstrate how the weighted profile concept can be applied to obtain a scoring function that enables ranking inhibitors by potency. In addition, the weights can be mapped onto residues thus providing a visually interpretable measure of binding on a residue-by-residue basis. Our approach is to use inhibitor potency data to extract an optimum set of relative weights for the interactions in the profile. The similarity between any interaction fingerprint and the weighted profile is measured by the Tanimoto similarity that in turn will depend on the weights assigned to each position in the profile. In this way the weights can be assigned such that they encode information about which interactions contribute most to potency. Though we realize that the potency of an inhibitor against a target protein is not solely determined by binding enthalpy, we assume here a direct correlation between potency and ligand−receptor interactions for simplicity, neglecting contributions from desolvation[4] and other entropic effects.

Strategy To Determine Weights. The weights in the weighted profile are assumed here to be non-negative and range between 0 and 1. We define the weights at positions that have a 1 in at least one of the SIFts to be positive, and the weights at positions that do not have a 1 in at least one of the SIFts to be zero (as illustrated in Figure 1). These weights will be determined using an approach such that the computed weights will represent the significance of each interaction in contributing toward overall protein−ligand binding. The approach followed here is a statistical learning method where the weights are learned using a training set such that the similarity between the weighted profile and SIFt is positively correlated with the inhibition potency. We will call this similarity (Tanimoto score) between the weighted profile and SIFt as the w-SIFt score. The reasoning behind the proposed methodology is as follows: the interactions that appear more frequently in high potent compounds are supposedly more important, so in order to boost the w-SIFt score of the high potent compounds the weights for those interactions will be calculated to be higher. Likewise, interactions that appear more frequently in less potent compounds are supposedly less important, and so in order to decrease the w-SIFt score of the less potent compounds these interactions' weights will be lower. Thus the overall weights in the weighted profile so determined will represent the importance associated with each SIFt interaction bit in the protein−ligand binding potency.

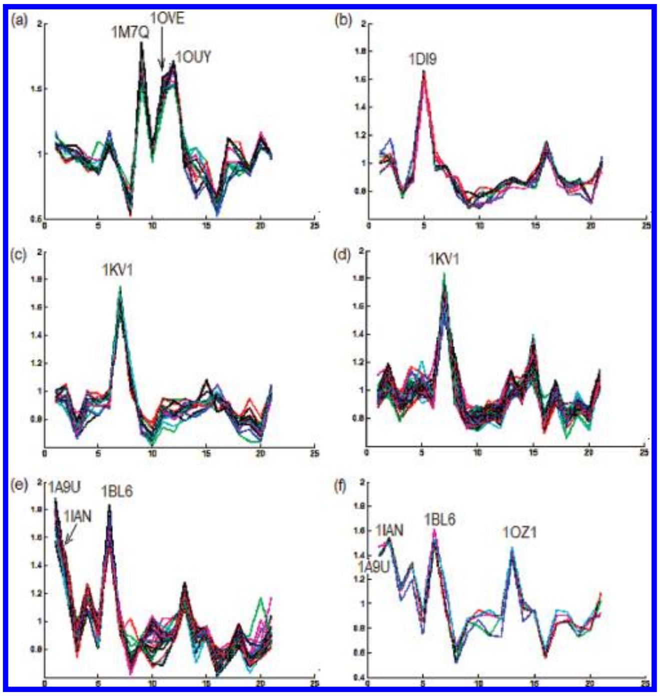

Figure 2. A sample of 6 clusters (out of 30 clusters) of ligand similarity profiles for P38α compounds. The clusters were based on Tanimoto similarity (defined by ROCS combo score) with the ligands of the 21 PDB complexes. Along the *x*-axis are the 21 PDB complexes and on the *y*-axis is the ROCS combo score.

Data Set Generation. The starting data set consisted of 675 P38α inhibitors whose potency (IC50) values have been reported in literature. The two-dimensional structure of these inhibitor compounds is also available. However, in order to generate SIFts we require the correct three-dimensional structure of the ligand binding mode to P38α. The three-dimensional binding poses of the ligands are obtained here by an indirect procedure where known PDB ligands are taken as reference to generate a pose that is subsequently optimized using a docking algorithm.

We have used 21 protein-small molecule crystal structures for P38α found in the PDB (having PDB codes 1A9U, 1BL6, 1BL7, 1BML, 1DI9, 1IAN, 1KV1, 1KV2, 1OVE, 1OUK, 1OUY, 1OZ1, 1M7Q, 1W7H, 1W82, 1W83, 1W84, 1YQJ, 1YW2, 1YWR, and 1ZZL). In order to identify a starting guess pose and optimal receptor structure for docking for each of the 675 inhibitors, we first search if any of the 675 compounds have a three-dimensional conformation that is structurally close to any of the 21 PDB ligands in their X-ray binding conformation. These guess structures are used to subsequently determine more accurate binding poses using docking.

An ensemble of three-dimensional conformations (on an average, 340 per molecule) was generated initially for each of the 675 compounds using Omega.[5] These conformations were queried using ROCS for structural similarity with the 21 PDB cocrystal ligands to generate a ligand similarity profile for each compound over the panel of 21 P38α ligands. Those compounds that did not have a similar 3D shape match and overlay to any of the 21 PDB ligands were filtered out. The similarity profiles for the remaining 362 compounds were then hierarchically clustered (into 30 clusters) based on their relative Tanimoto similarity. Figure 2 shows the clustered similarity profile for compounds in a sample of six out of the thirty clusters. The peaks in each cluster of ligand similarity profiles represent the structurally most similar PDB ligands for that cluster of compounds thus suggesting that the subset of compounds are very likely to bind to the receptor in a similar mode as the peak PDB ligand. For example in cluster 2(a) the subset of compounds is likely to bind with P38α similar to the ligands in 1M7Q, 1OVE, and 1OUY. The P38α structure(s) corresponding to the peak similarity in the cluster was thus chosen for docking the compounds represented in that cluster using Glide.[6] Cutoffs on both the docking score and the SIFt score between the final docked pose and the respective PDB ligand were used as criteria for final pose selection. In total, final poses were determined with confidence for 89 distinct p38 inhibitors.

METHODS

Objective Function for Determining Weights. Assume **s** represents a SIFt in a vector form and **w** represents the weighted profile. The w-SIFt score, let us call $\mathbf{T}_w$, is defined as the Tanimoto similarity between the SIFt and the profile. i.e.

$$T_w = \frac{\mathbf{s}\cdot\mathbf{w}}{\mathbf{s}\cdot\mathbf{s} + \mathbf{w}\cdot\mathbf{w} - \mathbf{s}\cdot\mathbf{w}}$$

The weights of the profile will be determined so as to obtain a w-SIFt score that correlates well with the experimentally determined potencies. We constrain the weights to be positive since in principle they represent the significance of the corresponding interactions. The objective of determining the weights can be mathematically stated as follows

$$\textit{To determine } \mathbf{w} \textit{ so that } T_w \propto -Log(IC50), \quad with\ w_i \geq 0$$

i.e. find **w** that corresponds to a straight line fit between T_w and $-Log(IC50)$ with highest correlation. The Pearson's correlation coefficient is chosen here to measure the extent of correlation. So, the objective function is formulated as

$$\underset{w_i \text{ s.t. } w_i \geq 0}{Maximize}\ CorrCoef(T_w, -Log(IC50)) \qquad (1)$$

where

$$CorrCoef(x, y) = \frac{\mathrm{cov}(x, y)}{\sqrt{\mathrm{var}(x)\cdot\mathrm{var}(y)}}$$

Since the objective function is complex and nonlinear, and the number of variables (i.e., weights) to alter is very large, we apply a stochastic optimization technique (Simulated Annealing[7]). The energy function for Simulated Annealing is defined here as the negative of the objective function defined in eq 1.

$$E_{SA} = -CorrCoef(T_w, -Log(IC50)) \qquad (2)$$

Handling Linear Dependencies in SIFts Data. The active site of P38α consists of 56 residues with each residue being represented by 10 bits,[8] making SIFt a 560 bit string. After eliminating the zero bits in the profile, there are 112 bits remaining implying that there are 112 weights to be determined, whereas there are only 89 SIFts from the docked structures described in the previous section. However not all the 112 nonzero bits in SIFt are independent of each other, since there could be co-occurrences (i.e., two bits simultaneously 'on' or 'off') and cross-occurrences (i.e., bits that are complementary to each other). There could also be additional statistically significant dependencies between bit pairs, i.e. two bit positions positively or negatively correlated for a significant percentage of the data. So we can use a dimensionality reduction technique to reduce if not eliminate these interdependencies and eventually compress the number of SIFt bits from 112 to a considerably smaller number without losing significant information. Thus, by doing so the number of weight parameters to be determined in the weighted profile is also significantly reduced. As the interdependencies in the SIFt are linear, we choose a linear dimensionality reduction technique for the data compression. The values of the SIFts in the reduced space need not be binary but have to be positive. We now only have as many weights to be determined as the dimension of the reduced space. After determining the weights in the lower dimensional space, the weights in the higher dimensional space (i.e., the original weights of the SIFts) can be obtained by an inverse transformation.

A linear dimensionality reduction technique involves a factorization of the original data matrix into two submatrices. Suppose the SIFt data set is represented as an $n \times m$ matrix, $\mathbf{S}^h$ where m is the number of SIFts and n is the length of a SIFt vector. During linear dimensionality reduction two submatrices **L**and $\mathbf{S}^l$ of size $n \times r$ and $r \times m$, respectively, are determined.

$$i.e.\ \underbrace{\mathbf{S}^h}_{n\times m} \approx \underbrace{\mathbf{L}}_{n\times r} \cdot \underbrace{\mathbf{S}^l}_{r\times m} \qquad \text{where } (n+m)r < nm$$

The **L** matrix represents the linear transformation from a higher n-dimensional space to a lower r-dimensional space. The matrix $\mathbf{S}^l$ represents the m SIFts in the lower dimensional space.

After performing the dimensionality reduction, the weights are first determined in the lower r-dimensional space and then back calculated in the higher n-dimensional space using the equation, $\mathbf{w}^h = \mathbf{L}\cdot\mathbf{w}^l$. Non-negative Matrix Factorization (NMF) is used here for dimensionality reduction of the SIFt space as the non-negative constraint imposed in this method helps to preserve the underlying physical interpretation of the weights.

Lee and Seung[9] demonstrated that NMF involves parts based learning of objects and is very effective and meaningful for dimensionality reduction in applications like image processing and text mining. NMF has been applied in several recent works in the context of computational biology and bioinformatics. Gao and Church[10] applied NMF as an unsupervised classification method for cancer identification based on gene expression data and found the method to be effective over other clustering techniques. Brunet et al.[11] have also used NMF on cancer related microarray data. The basis vectors in their work, called meta genes, represented distinct molecular patterns thus enabling them to extract meaningful

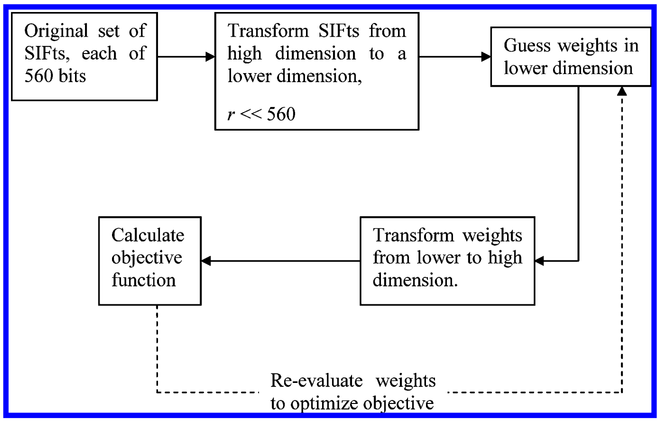
Original set of SIFts, each of 560 bits
Transform SIFts from high dimension to a lower dimension,
r << 560
Guess weights in lower dimension
Calculate objective function
Transform weights from lower to high dimension.
Re-evaluate weights to optimize objective

Figure 3. Illustration of the workflow involving dimensionality reduction and weights calculation.

biological information. In ref 12 Kim and Tidor used NMF on a large data set of genome-wide expression measurements of yeast and were able to detect local features in the expression space that mapped to functional cellular subsystems. Recently, Devarajan[13] provided a review of recent NMF applications in the context of biological informatics.

When NMF is applied to SIFts, the basis vectors represent underlying patterns of interactions between protein and the ligands as explained in the Results section. The original algorithm for NMF as described by Lee and Seung[9] has been used here for the dimensionality reduction. Figure 3 summarizes the workflow of the dimensionality reduction and the determination of the weights.

Determining the Dimensionality of the SIFt Space. The methodology described in the previous section involves a dimensionality reduction step that requires knowing *a priori* the dimensionality of the reduced space. Since we do not know the exact value of the reduced dimensionality in the case of SIFts, we build models based on some guess of the reduced space dimensionality using a training set and validate the models on a validation set. We use a 5-fold cross-validation procedure for training and validating the models here, as we do not have sufficient data to split into separate training and validation sets. The data set of 89 SIFts is divided into a training set (both for training and cross-validation) of 80 SIFts and a test set (for final testing) of 9 SIFts. In the cross-validation procedure, a model is built for each of the five training sets and is validated against its corresponding validation set. The validation error of an individual model is the sum of squared differences between the model prediction values and the experimental $-$ *Log*(*IC*50) values for the validation set. The overall cross-validation error of the five models is taken as the average of validation errors of the five individual models.

For each guess value of lower space dimensionality the overall cross-validation error is computed and that value corresponding to the least overall cross-validation error is accepted as the correct dimensionality of the lower space. Figure 4 summarizes the steps involved in determining the correct reduced dimensionality using the cross-validation procedure.

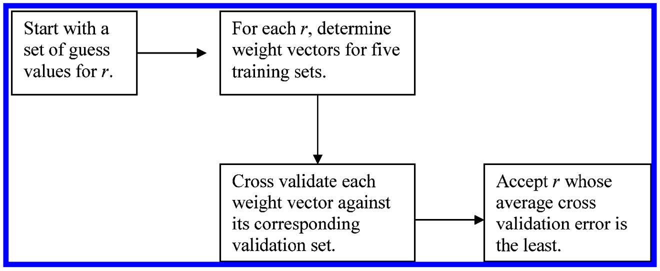
Start with a set of guess values for r.
For each r, determine weight vectors for five training sets.
Cross validate each weight vector against its corresponding validation set.
Accept r whose average cross validation error is the least.

Figure 4. Outline of steps to determine the correct dimensionality (*r*) of the reduced space.

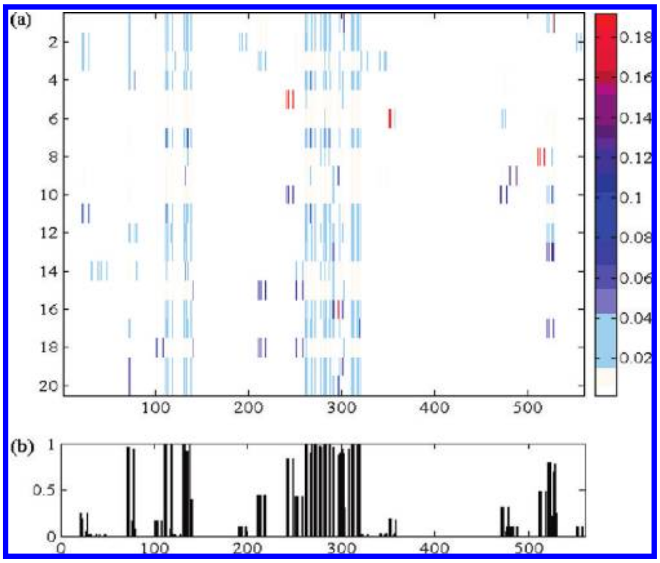

Figure 5. (a) Heatmap of the transformation matrix (**L**) from 560 bit-space to a lower dimensional space (of size 20). The panel on the right shows the numerical value range for the colors in the heatmap. (b) The average of all the SIFts in the entire data set (which is the profile defined in ref 2).

RESULTS

The cross-validation errors calculated for various guess values of dimensionality of the reduced space are shown in Table 1. The results show that a value of 20 for the lower space dimensionality corresponds to the least overall cross-validation error, implying that the given P38α SIFt data can be efficiently translated as a combination of 20 linearly independent vectors. Figure 5 shows a heat map representation of **L** which is a graphical illustration of the 20 basis vectors in terms of the original 560 bits. Each of these basis vectors represents an 'interaction pattern' which is a combination of individual interactions that were found to co-occur in the original SIFt data. Each entry in the basis vector corresponds to the importance of that particular bit in that pattern of interactions. Thus the basis vectors represent a meaningful combination of interactions due to the non-negative restriction on the elements of **L** matrix. Also, since the transformation matrix, **L**, is non-negative, we simply need to restrict our weights in the lower dimensional space to be positive in order to satisfy the criterion that the weights in the original 560 bit space should be non-negative.

The weight values of the weighted profile are provided as Supporting Information in Table S1. Figure 6 shows a graphical representation of the weighted profile(b), where each of the vertical bars correspond to the weight at that particular residue. Also shown in the figure is the unweighted profile(a) for the sake of comparison. In Figure 7(a), the w-SIFt scores of the training compounds, computed using the final weights model, are plotted against $-$*Log*(*IC*50) values. The SIFt training data are categorized into three classes (colored blue, yellow, and red in the figure) for better illustration and subsequent box plot analysis. The points in

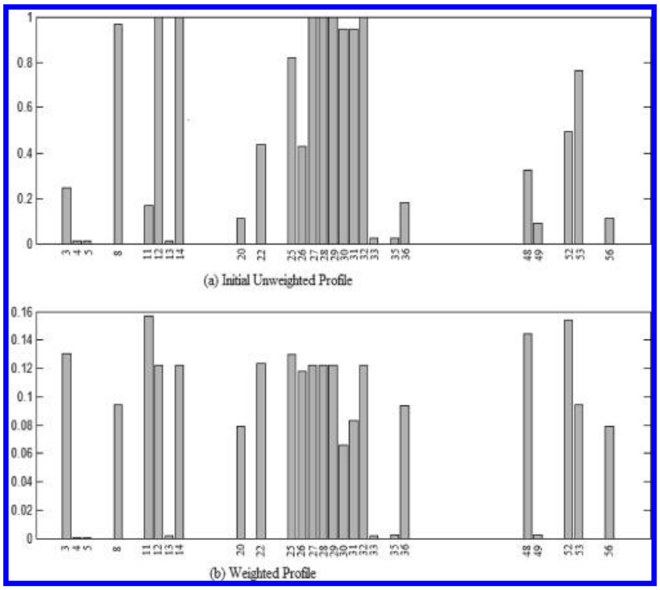

Figure 6. Comparison of the original unweighted profile (as defined in ref 2) and the weighted profile. For the sake of visual clarity, only the contact bit of the 10 bits at each binding residue is shown.

blue, yellow, and red correspond to highly potent, moderately potent, and least potent compounds, respectively. Figure 7(b) is the corresponding box plot representation showing the mean, quantiles, and outliers of the weighted profile scores (w-SIFt scores) for the three classes.

A similar analysis is done over the original unweighted profile scores. Figure 8(a) is a scatter plot of SIFt scores (i.e., scores computed with the unweighted profile and the initial definition of the profile, as defined in ref 2) against $-Log(IC50)$. The Pearson correlation coefficient between the unweighted profile scores and $-Log(IC50)$ values is −0.1046, while between the w-SIFt scores and $-Log(IC50)$ the Pearson correlation coefficient is 0.6040. The lack of correlation between the original profile scores and the $-Log(IC50)$ values is evident in the box plots shown in Figure 8(b).

The weighted profile model is tested for prediction accuracy on the test set of 9 SIFts kept aside before building the model (Figure 9). It is found that the test error is 2.6039 and is significantly close to the cross-validation error of 2.6364 (of Table 4.1), which is typically a good heuristic estimate of the model prediction error.[14]

The analysis done in Figure 7 was repeated for molecular weight and the docking score, in order to assess the performance of two null models. Figure S1(a) shows the scatter plot of the molecular weight against the $- Log(IC50)$ values, whereas Figure S1(b) is its corresponding box plot. Figure S2(a) is the scatter plot of −docking score against the $-Log(IC50)$ values with its respective box plot shown in Figure S2(b). The figures show that the molecular weight ($R = 0.2929$) and docking score ($R = 0.3415$) bear some correlation with the potencies though the deviation from the straight-line fit seems to be higher as evidenced in the respective box plots.

DISCUSSION

Typical physics based or empirical scoring functions are difficult to interpret: it is often not possible to extract information on what residues are driving potency and which interactions are more dispensable. The visual interpretation of the profile weights as illustrated in the previous section is perhaps the most powerful feature of the weighted interaction profiles described in this paper.

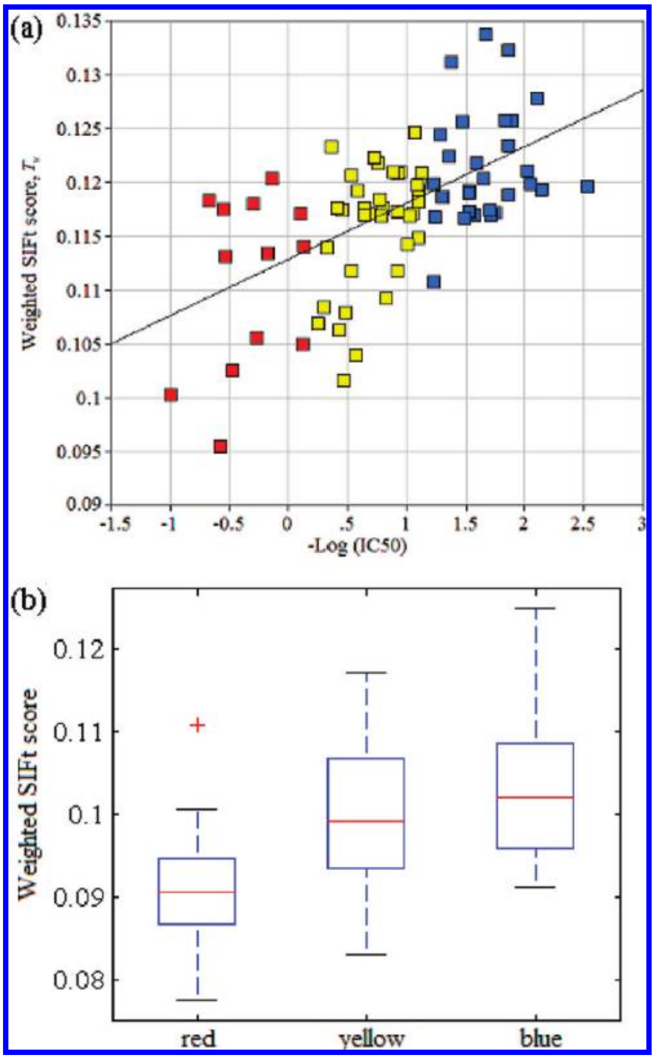

Figure 7. (a) Scatter plot of the weighted SIFt scores against $-Log(IC50)$. The points in blue, yellow, and red correspond to the most potent, moderately potent, and least potent compounds. The correlation coefficient $R = 0.6040$. (b) Box plots of the distribution of the weighted SIFt scores with respect to potency classes.

The binding pocket of P38α with a ligand bound to it (PDB 1BL7) is shown in Figure 10, with the key binding residues highlighted with purple, cyan, and white. It is observed that the weights illustrated in Figure 6(b) in fact reflect the relative importance of specific interactions in determining the potency of the P38α inhibitors considered in this study. In Figure 10, the most highly weighted residues are in purple; those with intermediate weight in cyan, and those least important for potency are in colored white. The majority of ATP competitive kinase inhibitors interact with the hinge region of the kinase via at least one hydrogen bond[2] mimicking the interactions made by the adenine moiety of ATP. In fact, these interactions are often used as constraints for filtering poses from docking experiments.[2,15] Not surprisingly, interactions with Met109, the key hydrogen-bonding residue in the hinge for P38α, are weighted heavily. In addition, Ala51 that makes hydrophobic contact with the typically heteroaromatic hinge binding sub-

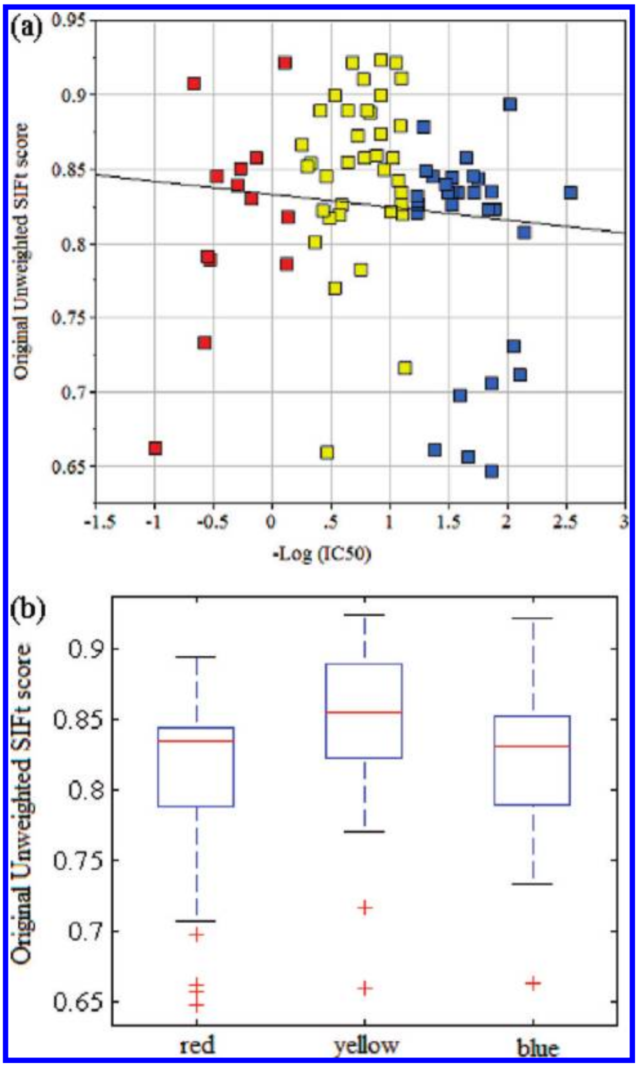

Figure 8. (a) Scatter plot of the original unweighted profile scores against $-Log(IC50)$. The points in blue, yellow, and red correspond to the most potent, moderately potent, and least potent compounds. The correlation coefficient $R = -0.1046$. (b) Box plots of the distribution of the unweighted SIFt scores with respect to potency classes.

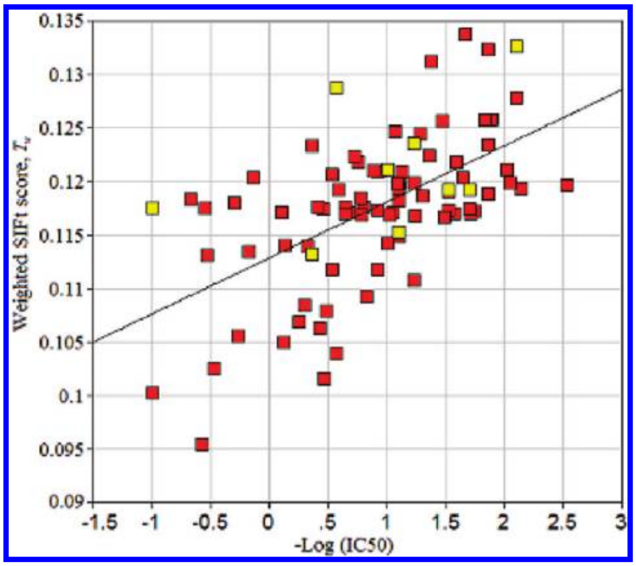

Figure 9. Scatter plot of the weighted profile scores against $-Log(IC50)$ for training (in red) and testing (in yellow) compounds.

Table 1. Cross-Validation Error of Models Built Using Different Values of the Lower Dimensionality in NMF

reduced dimensionality in NMF	5-fold cross-validation error of model
10	2.7433
15	2.8400
20	**2.6364**
25	2.8330
30	2.7329
35	2.8065
40	2.7162

stituents is identified as important for potency. Another nearly canonical interaction observed in the majority of kinase inhibitor cocrystal structures is with the conserved residue Lys53. In addition to these highly conserved interactions, the hydrophobic pocket and sugar cpocket regions of the ATP binding site received high weights. As is shown for example in Figure 11(a), inhibitors with substituents interacting with sugar pocket residues demonstrated increased potency over unsubstituted examples. Targeting the sugar pocket is a common strategy in kinase inhibitor design although it is not necessary to achieve potent activity in many kinases. The current analysis, however, indicates that this is an important region for p38α inhibition. In contrast, interaction with the P-loop of the kinase is not as important. Although many p38α inhibitors extend away from the hinge region and interact with Tyr35 (for example SB203580; PDB code 1A9U),[16] this interaction is not critical for potency. One possible rationalization may be that the P-loop in general, and Tyr35 specifically, demonstrates a high degree of flexibility in P38 and can adapt to a broad range of substituents. The P-loop can essentially clamp down on unsubstituted inhibitors like the imidazole example and open up when the imidazole is substituted as in SB203580. The hydrophobic (or selectivity) pocket shown in Figure 11(b) was the final region that was identified in our analysis as being critical for potency. The small Thr106 gatekeeper residue in P38α permits access to the hydrophobic pocket unlike in kinases with bulky gatekeeper residues, e.g., CDK2 (Phe) or Akt (Met). Many P38α inhibitors exploit this region with substituted phenyl groups that contact a cluster of hydrophobic residues lining the pocket. The weights determined from our analysis highlight the importance of these interactions for achieving potency against P38α. Finally, interactions with the hinge toward the solvent channel of P38α were in comparison much less important for potency. As substitution toward solvent is typically aimed at improving inhibitor solubility, physical properties, and selectivity,[17] it is not surprising that the weights determined from potency alone are not high. However, inhibitors with solvent channel substituents that made hydrophobic contacts with Val30 did receive relatively high weights in our analysis.

In addition to being interpretable, we have demonstrated that with an optimized set of target-specific weights, weighted profiles are able to rank order compounds based on potency. This is a marked improvement over our previously reported unweighted interaction profiles that are only able to provide enrichment for actives in virtual screening by filtering out binding modes that are inconsistent with known X-ray binding modes. The weighted SIFt scoring function could be used as a virtual screening tool for mining potent compounds from chemical databases. The first step of the virtual screening protocol would involve docking the inhibitors against the target protein and determining accurate poses based on a SIFt based

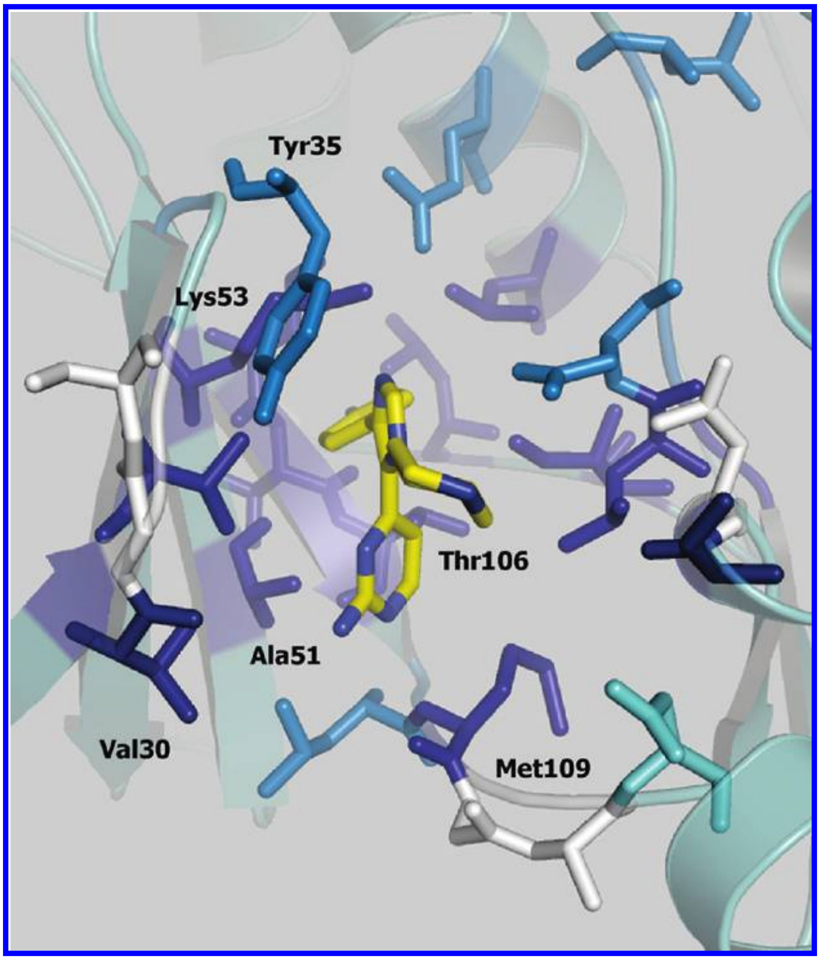

Figure 10. P38α with the key residues colored according to their weights. The residues in purple are the most highly weighted followed by residues in cyan, and the residues in white are the least weighted residues. Also shown in the figure are the labels for the residues referred to in the Discussion section.

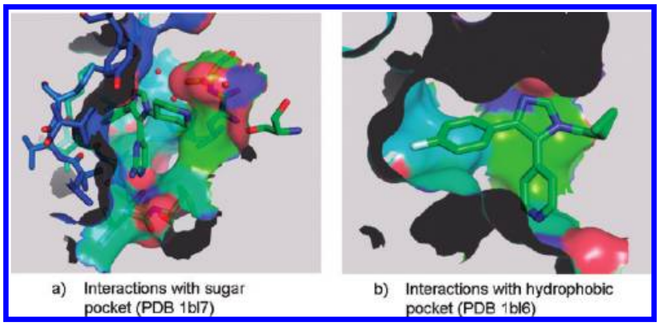
a) Interactions with sugar pocket (PDB 1bl7)
b) Interactions with hydrophobic pocket (PDB 1bl6)

Figure 11. The sugar pocket (a) and hydrophobic pocket (b) are identified from the w-SIFt analysis as important regions for potency.

filter as demonstrated by Deng et al.[1] The weighted profile and the SIFts of the docked poses are now used to compute the w-SIFt score, which is used as a ranking criterion.

This work presents a methodology for determining the weights on a relatively small set (89 compounds). The concept of weighting the bits in SIFt can be extended to determine other criteria such as selectivity of a compound toward two targets. Rather than training the weights for learning experimental potency values, the weights now have to be trained for learning the relative potencies expressed as Δ(pIC50) for example. The w-SIFt scoring function however suffers from the shortcoming that it is entirely based on assigning potency to protein−ligand binding interactions and does not include terms to delineate entropic contributions. There is however scope to combine the concept of weighting the interactions with other important ligand based terms like polar surface area, molecular weight, etc. that also play a critical role in protein−ligand binding.

ACKNOWLEDGMENT

We acknowledge contributions from Aldrin R. Denny and Aravind Prasad during the early stages of this work. We also thank Michael Corbley and Rainer Fuchs of Biogen Idec for helping us through the disclosure process.

Supporting Information Available: Table displaying weights of the SIFt contact only bits against the corresponding residues numbers and figures showing scatter and box plots of molecular weight and docking score against −Log(IC50). This material is available free of charge via the Internet at http://pubs.acs.org

REFERENCES AND NOTES

(1) 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.
(2) Chuaqui, C.; Deng, Z.; Singh, J. Interaction profiles of protein kinase-inhibitor complexes and their application to virtual screening. *J. Med. Chem.* **2005**, *48*, 121–133.
(3) Brewerton, S. C. The use of protein-ligand interaction fingerprints in docking. *Curr. Opin. Drug Discovery Dev.* **2008**, *11*, 356–364.
(4) Anderson, A. C. The process of structure-based drug design. *Chem. Biol.* **2003**, *10*, 787–797.

(5) *Omega, version 1.8.1*; OpenEye Scientific Software: Santa Fe, NM, 2004.

(6) *Glide, version 3.5*; Schrodinger Inc: New York, NY, 2005.

(7) Kirkpatrick, S.; Gelatt, C. D.; Vecchi, M. P. Optimization by Simulated Annealing. *Science* **1983**, *220*, 671–680.

(8) Singh, J.; Deng, Z.; Narale, G.; Chuaqui, C. Structural Interaction Fingerprints: A New Approach to Organizing, Mining, Analyzing, and Designing Protein-Small Molecule Complexes. *Chem. Biol. Drug Des.* **2006**, *67*, 5–12.

(9) Lee, D. D.; Seung, H. S. Learning the parts of objects by non-negative matrix factorization. *Nature* **1999**, *401*, 788–791.

(10) Gao, Y.; Church, G. Improving molecular cancer class discovery through sparse non-negative matrix factorization. *Bioinformatics* **2005**, *21*, 3970–3975.

(11) Brunet, J. P.; Tamayo, P.; Golub, T. R.; Mesirov, J. P. Metagenes and molecular pattern discovery using matrix factorization. *Proc. Natl. Acad. Sci.* **2004**, *101*, 4164–4169.

(12) Kim, P. M.; Tidor, B. Subsystem identification through dimensionality reduction of large-scale gene expression data. *Genome Res.* **2003**, *13*, 1706–1718.

(13) Devarajan, K. Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology. *PLoS Comput. Biol.* **2008**, *4*, n/a.

(14) Hastie, T.; Tibshirani, R.; Friedman, J. H. In *The elements of statistical learning*, 2nd ed.; Springer: New York, NY, 2003; pp 193−224.

(15) 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.

(16) 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.

(17) Fitzgerald, C. E.; Patel, S. B.; Becker, J. W.; Cameron, P. M.; Zaller, D.; Pikounis, V. B.; O'Keefee, S. J.; Scapin, G. Structural basis for p38 alpha MAP kinase quinazolinone and pyridol-pyrimidine inhibitor specificity. *Nat. Struct. Biol.* **2003**, *10*, 764–769.

CI800466N