Published in final edited form as:

Proteins. 2011 September ; 79(9): 2746–2763. doi:10.1002/prot.23103.

Surface-Based Protein Binding Pocket Similarity

Russell Spitzer, Ann E. Cleves, and Ajay N. Jain*

University of California, San Francisco, Dept. of Bioengineering and Therapeutic Sciences, Helen Diller Family Comprehensive Cancer Center

Abstract

Protein similarity comparisons may be made on a local or global basis and may consider sequence information or differing levels of structural information. We present a local 3D method that compares protein binding site surfaces in full atomic detail. The approach is based on the morphological similarity method which has been widely applied for global comparison of *small* molecules. We apply the method to all-by-all comparisons two sets of human protein kinases, a very diverse set of ATP-bound proteins from multiple species, and three heterogeneous benchmark protein binding site data sets. Cases of disagreement between sequence-based similarity and binding site similarity yield informative examples. Where sequence similarity is very low, high pocket similarity can reliably identify important binding motifs. Where sequence similarity is very high, significant differences in pocket similarity are related to ligand binding specificity and similarity. Local protein binding pocket similarity provides qualitatively complementary information to other approaches, and it can yield quantitative information in support of functional annotation.

Introduction

Comparisons of small molecules based on surface characteristics including both shape and polarity have been shown to yield separable similarities for pairs of molecules that bind the same protein sites from those that do not.¹ Variations of the approach have been used in virtual screening² and for identifying molecular superimpositions for use in constructing binding site models for affinity prediction based on the structures and activities of small molecules targeting a single protein cavity.^{3,4} The approach is based on defining molecular observers around two aligned molecules and comparing what they “see” in terms of distances to the molecular surfaces and the polarity of those surfaces (efficient solutions exist for the problem of identifying the optimal alignment and conformation of one molecule onto another). Recently, we have extended the approach to comparisons of *concave* surfaces such as protein binding pockets.⁵

This extension (from global/convex to local/concave) required two significant additions to the similarity computation: 1) a method to define the spatial scope of the desired comparison; and 2) methods to avoid degeneracies such as solvent accessibility that are present in the concavity-comparison case and do not arise otherwise. The software implementing the algorithms for pocket construction and ligand activity prediction constitute a new module within the Surflex platform, called Surflex-PSIM (**P**rotein **S**imilarity). In this paper, we present the details of the similarity computation and optimization algorithm and application of the approach to three sets of related protein structures of varying diversity, two consisting of sets of protein kinases and the other

*Correspondence and Proofs to Ajay N. Jain: **US Mail:** Prof. Ajay N. Jain, University of California, San Francisco, 1450 3rd Street, MC 0128, PO Box 589001, San Francisco, CA 94158-9001, **Federal Express:** Prof. Ajay N. Jain, University of California, San Francisco, 1450 3rd Street, Room D373, San Francisco, CA 94158, Voice: (415) 502-7242, Fax: (650) 240-1781, ajain@jainlab.org.

consisting of evolutionarily-divergent ATP binding proteins. We compare the results obtained from these local structural comparisons to those obtained based on sequence comparisons. Whereas the sequence-based approach is clearly able to identify ancestral relationships between proteins, the surface-based approach offers complementary information allowing for more subtle distinctions that relate to protein function, especially those functions related to non-covalent ligand binding within protein cavities. We also present direct comparisons to other methods on three sets of heterogeneous protein structures reported by Kahraman et al.,⁶ Hoffman et al.,⁷ and Yeturu et al.⁸

The first kinase set consisted of 45 structures of the MAP-kinase family EC 2.7.11.24. The second kinase set contained 183 protein structures, corresponding to 26 different human kinases, for which ligand binding data were available, as in Fabien et al.⁹ Our quantitative pocket comparisons mirrored those found by Kinnings et al.¹⁰ This separation efficiency was superior to using sequence-based methodologies for several of our target ligands. The kinases represented a group of proteins that diverged relatively late in evolutionary time (based on amino acid similarity), whose basic function is the same but with differences in specificity and in regulation. Differences among these proteins included very subtle alterations of ligand binding pockets, and even the more substantial changes still preserved overall protein architecture and clear global sequence similarities.

ATP has been a metabolically important small molecule for the entire duration of evolution, so those proteins that make use of its properties have evolved a number of different structural mechanisms for doing so over a very long time span. The ATP-bound set contained 267 protein structures, corresponding to 120 annotated Enzyme Commission numbers, and many different families of proteins (including 6 helicases, 12 ligases, 39 metabolic kinases, 28 protein kinases, 13 polymerases, 5 protein folding chaperones, 29 tRNA synthetases, 8 transcription factors, and 21 transporters) from 78 species. While all of these proteins bind ATP, there is a great deal of variety in both the sequences of the proteins as well as the physical motifs used for binding. Comparisons using pocket similarity here identified cases as in the kinase set with high sequence similarity and a range of pocket similarities. More interestingly, a number of cases were identified with very high pocket similarity where the sequence similarity was so minimal as to be undetectable using standard sequence alignment methods.

Analysis of the data sets of Kahraman (100 proteins binding 9 different ligand types), Hoffman (100 proteins binding 10 different ligands), and Yeturu (26 protein structures, with 51 ligand binding sites for 4 ligands) illuminated the differences between the Surflex-PSIM approach and other methods. The Kahraman set includes protein cavities of widely varying volume, with the Hoffman set designed specifically to avoid such gross heterogeneity. In these cases, Surflex-PSIM performed statistically indistinguishably from the best of previously reported methods. In contrast to those methods, the PSIM approach showed only a limited correlation with pocket volume differences, while showing a significant relationship between pocket similarity and cognate ligand similarity. On the Yeturu set, whose focus was on the binding sites of four ligands, where each binding site was represented by highly similar variants, the Surflex-PSIM approach yielded a perfect segregation of the sites by cognate ligand.

Methods and Data

Data sets comprising human protein kinases and evolutionarily diverse ATP-bound proteins were the subject of our primary analyses of proteins with related functions or related ligands. Comparative analyses were also carried out on three sets of proteins with diverse functions and ligands. The following describes the details of these sets, then the computational

similarity methods, and finally the statistical analysis approaches. Data and computational protocols are available for download (see <http://www.jainlab.org> for details).

Molecular Data Sets

Related Proteins—Two sets of kinases were used in this study. One was the set of 45 protein-ligand structures curated in the BindingMOAD database¹¹ that corresponded to EC number 2.7.11.24, of which 39 were human proteins, 4 mouse, and 2 rat. The second kinase set was obtained from a previous work on binding site comparisons by Kinnings and Jackson.¹⁰ The study analyzed 351 structures of 76 different kinases, of which 316 structures spanning 64 different kinases had bound ligands, and of those we employed 183 structures representing 26 human kinases for which binding affinities were available. We focused on the set of structures containing bound ligands in order to simplify specification of binding site location. Structures for the ATP data set were obtained from the RCSB PDB database.¹² The structures were obtained from a query for all structures, which contained a ligand identified as ATP with a sequence similarity filter of 95%. This resulted in 267 protein crystal structures (36 Archaeal, 125 Bacterial, 94 Eukaryotic, 10 Viral, and 2 synthetic), which were inspected manually to ensure that in every structure ATP was inside a binding site. Of note, the set was dominated by proteins with very low sequence similarity, with 95% of all protein pairs from the set having less than 20% sequence similarity, assessed by global sequence alignment by Needleman-Wunsch.

Diverse Proteins—Three sets of heterogeneous protein structures reported by Kahraman et al.,⁶ Hoffman et al.,⁷ and Yeturu et al.⁸ were used to make direct comparisons between PSIM and other methods. The Kahraman Set considered 100 protein structures, comprising multiple subsets of sequence-dissimilar proteins that each bound the same ligand type, as follows: PO4 (20 structures), Heme (16), NAD (15), ATP (14), FAD (10), AMP (9), FMN (6), glucose (5) and sex hormones (5). This set was characterized by significant diversity in both ligand size and in corresponding overall binding pocket volumes. The Hoffman Set was designed specifically to mimic the Kahraman set, but to limit the effects of diverse ligand sizes and pocket volumes on the binding pocket comparisons. It consisted of 100 protein structures, with 10 examples for each of 10 ligands (PDB ligand codes follow): 1PE, BOG, GSH, LDA, LLP, PLM, PMP, SAM, SUC, and U5P. The Yeturu Set included multiple binding sites for each of four ligands (methotrexate, indinavir, citrate, and phosphoglycolic acid), but the alternate binding sites were generally highly similar, including, for example, symmetry-related sites within a single protein structure.

Computational Methods

The basic notion behind our approach to molecular comparison is that we ought to make comparisons of molecules based on what their binding partners “see.” For ligands, we want to compare them based on the surfaces moieties that can be recognized by proteins. Given an alignment of two molecules, we define a similarity function that compares distances to the molecular surfaces from observer points surrounding the molecules. Computing the similarity requires identification of the alignment that maximizes this function. The observers are placed on a uniform grid of points with spacing λ . The points are weighted based on their minimum distance to the molecular surface, retaining a set of observers that correspond closely to a chosen distance γ (sharpness is controlled by σ). This identifies a finite set of observer points that are all “outside” the molecules. Where molecular surfaces are largely congruent in terms of both shape and polarity, the observer points will “see” the same things in the optimal alignment between the molecules.

Figure 1 illustrates the concept. In the case of small molecule ligands, we use 2.0Å, 4.0Å, and 0.2 for λ , γ , and σ , respectively. The similarity function itself is a normalized sum of

Gaussian functions of the differences in distance from each observer point to each molecule's surface. Such differences are computed for the minimum distance to any surface point (which gives the molecular shape), the minimum distance to a donor surface or formally positive atomic surface, and the minimum distance to an acceptor or negatively charged surface. Directionality and charge magnitude are also taken into account. Details can be found in the original paper.¹ The overall effect is that following alignment optimization, for molecules that can exhibit very similar molecular surfaces, both in terms of shape and disposition of polarity, the similarity function will return a value close to 1 whether or not the underlying molecular scaffolding is similar. The function itself is continuous and piecewise differentiable, which makes it suitable for computational optimization. For small molecules, where both conformational flexibility and relative alignment must be optimized, the procedure involves a divide-and-conquer strategy to address the conformational problem, heuristic search approaches to address the gross alignment, and local gradient-based optimization for final pose refinement.^{1,2}

As seen in Figure 1, the choices for λ , γ , and σ yield sensible results for ligands, where we wish to compare the outside of one ligand to the outside of another. For proteins, these values do not typically lead to sensible characterization of a specific pocket. At right in Figure 1, the mapping to protein site comparison is shown. By choosing a tighter grid ($\lambda = 0.5$), at a closer spacing ($\gamma = 0.5$), with a thinner shell of highly weighted points ($\sigma = 0.02$), within a specified radius of a point within the pocket in question, we can achieve the desired behavior: local comparison of concavities. For the experiments in this paper the point used is the centroid of a co-crystallized ligand, but ligands are not required and automated pocket detection could just as easily designate an approximate pocket center. For protein similarity computations, conformational variation is not explored, and the alignment optimization is carried out through sampling orientations quite densely to identify reasonable starting points for gradient-based local optimization of the similarity function.

Figure 2 shows how the approach is applied to two divergent human kinases (CDK2 and c-MET), which share less than 20% sequence identity but nonetheless have common ligands such as staurosporine. These proteins were brought into alignment by maximizing our local pocket similarity metric. Local sequence changes (proline and tyrosine replaced with glutamine and phenylalanine) yielded relatively little difference in the surfaces presented by these pockets. In Figure 3, a visualization of the similarity comparison is presented. The left image shows the placement of the observation points from the pocket alignment (superimposed around the transformed ligands) while the image on the right visualizes the similarities observed from these points. Red sticks represent similar negatively charged surfaces, blue represent similarities in positive surfaces, and green sticks correspond to similarities in hydrophobic surfaces. It is the kinase hinge region responsible for binding ATP that contains the most congruent parts of both pockets.

Computational Procedures

Detailed scripts for generating the results presented here are available in the data archive associated with this paper. Briefly, the procedures included automatic conversion of primary PDB files into mol2 format in order to address bond-order and protonation for both proteins and extracted ligands. All-by-all comparisons of proteins with pocket locations identified by ligand binding sites was performed automatically using the procedure outlined above. Such comparisons yielded similarity scores and alignment transforms among each pair of protein pockets. These were used to build fully aligned trees of protein structures using a greedy approach, adding proteins to a growing tree by seeking the next highest similarity for a protein outside the tree, using the proper pairwise transform combinations to bring all proteins into mutual alignment. Care was taken to define binding site locations and scope

using uniform methods to allow for automatic induction of protein alignment trees from initially unaligned proteins.

In the comparison of proteins of closely related function, we employed parameters for the similarity computation of $\lambda = 0.5$, $\gamma = 0.5$, and $\alpha = 0.02$, with even weighting of hydrophobic and polar features. We made use of a single definition of binding site scope from the final joint mutual alignment of all proteins. This was done for both kinase sets and for the ATP set. For comparing pairs of proteins from sets of widely divergent character, we employed parameters for the similarity computation of $\lambda = 0.5$, $\gamma = 1.0$, $\alpha = 0.02$, and a 0.5 weighting of polar features relative to hydrophobic features. The binding site scope was the union of scopes from each protein site within a given pair. Binding site scope for each protein site was focused on the interaction zones between ligand and protein. This definition resulted in only a weak relationship to pocket volumes, focusing instead on local chemical surface characteristics.

In what follows, Surflex-PSIM will be abbreviated as PSIM for the sake of brevity.

Results and Discussion

We considered three data sets from closely related proteins and three sets from heterogeneous proteins and will discuss results for the two classes in sequence. Among related proteins, we applied the PSIM computation to three different levels of protein structural diversity. The most closely related protein structure set was derived from BindMOAD, containing all protein-ligand complexes with EC number 2.7.11.24 (mitogen activated protein kinases). This comprised 45 structures of three different kinases (p38 α , JNK3, and ERK2). We then considered a larger and more diverse set of human protein kinases based on the work of Kinnings et al., composed of 316 structures of 64 different protein kinases. Last, we considered a set of 267 protein structures, all bound to ATP, but spanning highly divergent protein families (e.g. ATP-binding domains of ABC transporters and DNA/RNA polymerases).

MAP Kinases

The set of 45 MAP kinase structures included 3 gene products: 29 MAPK14 (p38 α), 8 MAPK1 (ERK2), and 8 MAPK10 (JNK3). We computed an all-by-all comparison of these structures using PSIM, and Figure 4 shows the resulting tree of similarity. The different kinase families were segregated well, with distinctions also made between different species and mutant proteins. The most similar pair (1WBT and 1WBS) contained wild type human p38 α bound to nearly identical ligands, differing only by a carbon/nitrogen swap in a heterocycle.

Following the nodes away from the root of the tree, we continue to see a substructure of the first two ligands bound to 1WBV, with excursions from the 1WBT ligand envelope occurring as we move up the tree. The ligand of 1OVE makes two separate protrusions from the binding envelope of the 1WBT ligand. Pocket conformations of 1WBS, 1WBV, and 1OVE are shown relative to 1WBT at left in Figure 4. The very different (and rigid) ligand of 1OVE binds to a different DFG configuration of the kinase altogether (the inactive conformational form). However, the structure is still correctly grouped with the p38 α exemplars, and it is correctly aligned with respect to the common hinge-binding elements shared by all of the ligands.

To formalize the relationship between the taxonomy represented by the tree in Figure 4 and ligand structures, we computed similarities for different groups of p38 ligands based on the tree structure. The set of protein conformational variants rooted at 1WBT and including

those proteins up to depth 4 (7 structures total) defined a group of ligands that were bound to highly similar pockets. The set of p38 variants at depth 10 or greater (12 structures, at the same level as or above 1OVE in Figure 4) defined a group of ligands bound to dissimilar pockets from those near 1WBT. Pairwise 3D ligand similarities among the pocket group near 1WBT were higher than those between that group and those distant from 1WBT (ROC area 0.77, $p \ll 0.01$ by resampling). Considering the ranked list of ligand similarities, pairs at the top of the list were over 20-fold more likely to come from proteins belonging to the group near 1WBT than to the cross-pairs that included a ligand from near the root and one distant from it. This will be further quantified below in comparison to other methods.

The resulting alignments have two desirable properties. First, similarity between protein pocket variants of the same flexible enzyme is related to the similarity of ligands that bind the pocket variants. Second, alignments that optimize local surface similarity preserve the geometry of parts of the protein surface that remain congruent when other parts of the protein binding pockets change significantly. The second property stems from the definition of the similarity metric as one that rewards similarity as opposed to penalizing differences.

Kinase Ligand Binding Profiles: Kinnings Data Set

One of the main goals of the structural analysis of proteins is the ability to differentiate proteins based upon their ligand binding preferences. The 183 structure/26 protein human kinase data set presents a particularly appealing target for differentiating ligand binding because of the difficulty of designing inhibitors with high specificity and the value of designing inhibitors with such specificity for cancer research. The work done in Kinnings et al.¹⁰ utilized enrichment testing for judging the success of their similarity metric as a methodology for determining the binding preference of proteins.

Given a query ligand Y, the idea was to identify proteins that would also bind Y by comparing their structures to the structure of a protein bound to Y. Enrichment was calculated based on the number of structures in the top 5 percent of the ranked list whose corresponding protein was inhibited the target ligand with a K_i of less than 10 μM . The enrichment score was defined as $((Ah/Th)/(A/T))$: Ah was the number of structures which bind the target ligand in the top 5%; Th , the total number of structures in the top 5%; A , the total number structures which bind the target ligand; and T , the total number of structures. P-values for enrichment scores were computed using a hyper geometric distribution. Computing enrichment in this fashion, PSIM yielded nearly identical results to those of Kinnings (Table 1).

However, it is important to understand that the results for a number of cases are dominated by the presence of multiple alternative protein structures for the cognate protein of the query ligand. The test considered all protein structures as being separate individuals, so even those structures that represented alternate conformations of the *same protein* were considered as being distinct in the enrichment testing. So, given a structure Z of ligand Y bound to protein X, an alternate conformation of protein X corresponding to structure Z' would positively influence the enrichment associated with ligand Y. Enrichment scores computed in this fashion can be somewhat misleading. For example the target structure for the ligand SB203580 is 1A9U, which is a crystal structure of p38 α bound to SB203580. The top 5% of structures similar to 1A9U are *also* crystal structures of p38 α , entirely dominating the enrichment computation (p38 α has 34 representative structures in the data set).

This bias is not apparent in all of the ligands tested. For example, results for Tarceva showed a significant enrichment factor that was obtained by the high ranking of structures from several different proteins. The target structure for the enrichment of Tarceva was 1M17 (EGFR bound to Tarceva). The top 5% most similar structures contained structures of the

proteins SRC and LCK. LCK has one of the most similar sequences to EGFR of the proteins in the set, but it was ranked below SRC, which has much less sequence similarity.

We made an alternative analysis that removed the bias of multiple structural exemplars by defining the similarity of two proteins as the maximum similarity obtained using the PSIM method computed over all pairs of structures of the two proteins. We employed Receiver Operator Characteristics (ROC) analysis to determine whether pairs of proteins which both bind the same ligand had significantly higher similarity than protein pairs where one binds a particular ligand and the other does not. Resampling was used to determine the significance of the ROC areas. In 6/10 cases, the ROC area exhibited significant separation ($p < .05$), showing higher similarity among protein pairs that shared ligands. The ROC plot for Tarceva is shown in Figure 5 (ROC area 0.77, $p < .002$). The alignment of EGFR with SRC gives a sense of the high local surface similarity that drives the high score (green arrows) along with significant differences that reflect the relatively low sequence similarity of the two proteins (red arrows).

The combination of sequence divergence coupled with surface epitope preservation is a common theme in biology. As sequence divergence increases, such 3D structural motifs are detectable with a similarity metric that focuses on surfaces and ignores sequence information.

Motif Discovery: ATP-Binding Data Set

The detection of motifs in proteins was originally and is still primarily done through the identification of sequence-based motifs or the computation of sequence similarity to known protein domains. These motifs are then used for the annotation and classification of unknown proteins. While these methods are extremely effective for long sequence motifs and proteins that share significant evolutionary history, they have limited ability to detect short discontinuous motifs and protein similarities based on convergent structures rather than amino acid homology from shared ancestry. Utilizing a three-dimensional structure-based method allows for the discovery of motifs that are discontinuous and may reduce false attribution when sequences are similar but a small change has significantly altered the structure.

ATP is an attractive ligand for considering distantly related proteins because of the long history of the molecule with respect to the evolution of life. Protein motifs that bind ATP have been projected to be among the very first to have evolved.¹³ This timeframe allowed for the development of many parallel ways of utilizing the molecule and also significant time for convergent pathways to produce congruent structures. Such common structural motifs can be discovered using a method analogous to the discovery of protein families with sequence similarity. Instead of grouping proteins together based upon common sequence and extracting a motif, proteins can be clustered based upon common binding site structure to reveal common structural motifs.

The 267 ATP-bound structures from the PDB were a particularly diverse set, including 36 Archaeal, 125 Bacterial, 94 Eukaryotic, 10 Viral, and 2 synthetic proteins, with 95% of the pairwise sequence similarities being less than 20%. Using PSIM, clusters were created from the ATP data set by first forming a fully connected graph with protein structures as nodes and PSIM similarity values as the weights on edges. The edges were then filtered to only retain highly significant ($p < 0.001$) edges based on a similarity threshold derived from a set of ~100,000 randomly selected unrelated protein pair similarities. The resulting clusters were then annotated based upon known protein functions. Figure 6 depicts the overall cluster and highlights subclusters of particular interest.

Kinases—One of the prominent clusters contained many of the structures from the work discussed previously. Kinases have been well annotated both for functional sequences and structural motifs. Sequence motifs have been defined for various regions of the binding domain and have previously been used to classify unknown proteins as kinases. Without the aid of this prior information, based purely on local binding site surface similarity, PSIM separated the kinases as a distinct cluster (Figure 6, lower right). The cluster also included the Pseudo-Kinase STRAD α , which contains a known kinase sub-domain but lacks some required catalytic residues, suggesting that the methodology is clustering based on binding modality and not catalytic activity. When the kinases are put into a common alignment induced from their surface similarity, a pronounced pocket is visible (Figure 7) showing the consensus structure used by these kinases to bind ATP. The most conserved portions of this surface are located in the nucleotide-binding region while greater variability exists near the phosphate tail. The similarity in all of the kinase binding sites stems from the common binding mode used by these proteins in their interactions with ATP.

However, different protein families make use of different “tricks” in binding ATP. Figure 8 highlights the binding modes of ATP within three different clusters from the global clustering. For kinases (lower right), disposition of the nucleotide head is largely conserved, mirroring similarities in binding the hinge region of kinases (so-called since it serves as a “hinge” between two protein domains, as seen in Figure 2). The phosphate tail exhibits greater variability, apparently reflecting the differing specificity of kinases with respect to transfer of phosphate to substrates. By contrast, the GKST loop proteins are strongly conserved in the phosphate tail region with great variability in the nucleotide portion, and the TRNA Synthetases show a more closely conserved ATP binding mode overall that is different from both. Note that since PSIM rewards surface congruence (as opposed to minimizing overall deviations), those regions of the binding sites that are most similar can be recognized without being disrupted by the more variable regions.

Although many of the known sequence features for kinases are short and discontinuous,¹⁴ a proper structural alignment should overlay the common residues making sequence based motif discovery more tractable. Such an alignment also guarantees that the residues not only share similar placement in sequence space but also share a similar relative structural location and enhances the chance that function is also shared. Extraction of sequence-based motifs from sequence alignments derived from the PSIM aligned structures produced many well-known features of kinases. Sequence-based motifs were analyzed using UCSF Chimera¹⁵ on the PSIM kinase pocket alignment. Motifs representing the DFG activation loop and the invariant lysine implicated in catalytic rate were found as well as variations on the well-known hinge-binding motifs¹⁶ (Figure 9). These small motifs would be nearly undetectable in a sequence based approach applied to as diverse a group of proteins as our ATP set. Motifs such as GXGXXG would have been particularly difficult due to the lack of specificity but a structure shaped by this motif in these protein contexts provides ample specificity to be discovered with the assistance of a 3D method.

GKST Loop—One of the oldest motifs known (both historically and evolutionarily) is the P-loop-containing triphosphate hydrolase fold,¹³ which has been noted for its conserved GKST motif. The loop is also known as the Walker-A motif. This short and weakly defined sequence motif (GXXXXGK[TS]) would be difficult to extract from a set of proteins as diverse as the ATP set based on sequences alone. We observed, however, that the corresponding structures yield an extremely well defined binding motif for ATP. The largest cluster (Figures 6 and 8, green cluster) that PSIM generated corresponded to this motif and exhibited a remarkably closely conserved tail conformation for ATP (Figure 8). The PSIM clustering correctly grouped proteins with widely varying evolutionary history. Figure 10 shows how the GKST structural loop binds the phosphate tail of ATP in conjunction with a

magnesium ion. All of the phosphate tails exhibit a nearly identical orientation with the nucleotide head free to be handled differently depending on the protein involved.

The remarkable structural conservation comes despite very significant sequence variation. Figure 11 shows the difference between the sequence alignment derived from the PSIM structural alignment and that derived by ClustalW¹⁷ purely based on sequence. GKST stands out as a significantly conserved motif but we also gain insight into some of the more drastic changes that can occur within this motif. In some cases, ClustalW was able to correctly identify the sequence motif, but in many cases, the motif was too weak. This is most likely due to the shortness of the GKST motif and the presence of many other similar regions in the proteins in question. The example of PDB structure 3FKQ from E. Rectale is particularly striking, since it is the only protein lacking the highly conserved lysine within the entire set. In this case the lysine was replaced with a threonine but ATP still binds, and does so in a manner nearly identical to those proteins bearing the canonical motif. ClustalW yields an unrelated sequence for 3FKQ in the multiple alignment, but the pocket similarity is sufficiently strong ($p < 0.001$) that PSIM was able to both place the protein in the correct cluster as well as identify the correct alignment correspondence with the other proteins in the family.

Heterogeneous Protein Data Sets

Among heterogeneous proteins, we applied PSIM to three different data sets (see Methods and Data for details), each constructed by different research groups reporting new protein pocket similarity approaches, and with each addressing slightly different considerations. The Kahraman Set considered 100 protein structures, comprising multiple subsets of sequence-dissimilar proteins that each bound the same ligand (e.g. ATP, AMP, PO₄, and NAD), with significant diversity present in both ligand size and binding pocket volume. The Hoffman Set was designed specifically to mimic the Kahraman Set, but to limit the effects of diverse ligand sizes and pocket volumes on the binding pocket comparisons. The Yeturu Set included multiple binding sites for each of four ligands, but the alternate binding sites were highly similar, including, for example, symmetry-related sites within a single protein structure (e.g. PDB code 1EGH, with 6 symmetric phosphoglycolic acid binding sites).

Kahraman Set—Results from application of a spherical harmonic shape method (the “cleft interact” version) due to Kahraman et al.⁶ and from an atom-cloud comparison method (sup-CK) due to Hoffman et al.⁷ measured the ability of a method to rank pockets that bound a particular ligand as being similar to a protein pocket that also bound that ligand. Results were reported using standard ROC area analysis, with the average of areas resulting from 100 different rankings (one each for each protein in the data set). The spherical harmonic approach yielded a mean ROC area of 0.77 (no standard deviation was given).⁶ The sup-CK method,⁷ using parameters tuned for the Kahraman Set, yielded 0.86 ± 0.14 . The PSIM approach yielded 0.79 ± 0.19 . None of these methods exhibited differentiable performance from one another, and none performed better than volume alone, which yielded 0.88 ± 0.14 .⁷ For these proteins, average pocket volume was strongly correlated with ligand size, with PO₄ having the smallest binding sites ($445 \pm 118 \text{ \AA}^3$) and FAD having the largest ($2099 \pm 224 \text{ \AA}^3$).⁶ Of the three methods, the PSIM approach was least strongly related to volume, with site comparisons being focused on observations of local chemical surface similarities and the binding pocket scope being limited in this computation to the portion of the binding pockets proximal to their cognate ligands.

Hoffman HD Set—Given the confusing influence of volume on interpretation of results, the sup-CK authors produced a homogeneous data set: ten binding sites for each of ten ligands, but where both the ligand and pocket sizes had far less variance than in the

Kahraman Set.⁷ On this set, volume alone yielded a mean ROC area of 0.65 ± 0.15 , sequence similarity 0.58 ± 0.09 , and sup-CK and sup-CKL each yielding *tuned* performance of 0.71 ± 0.19 and 0.75 ± 0.16 , respectively.⁷ PSIM yielded 0.76 ± 0.15 using precisely the same parameters as used for the Kahraman Set. While the sup-CK method showed a substantial performance decrease when volume was factored out of the comparisons, the PSIM approach yielded nearly identical performance on both sets.

Yeturu Set—Quite a different question was addressed by one of the data sets in the study by Yeturu et al.⁸ in which the PocketMatch algorithm was introduced. For this set, four different ligands (citrate, indinavir, phosphoglycolic acid, and methotrexate) in 26 total crystal structures (Figure 3 from the paper) yielded 51 binding sites, many of which were multiple minor variants within a single crystal structure. Given that the PSIM approach can make fine distinctions among local pocket differences, sensitivity to natural local variation may have posed some difficulty. However, as shown in Figure 12, PSIM showed perfect segregation of the binding sites based on ligand type (using exactly the parameters as used for the Kahraman and Hoffman Sets).

The methotrexate and indinavir cases generally represented very minor variations on nearly identical binding sites. In those cases, similarity scores were nearly all above 9.0. For phosphoglycolic acid, 6/11 sites were from a symmetric arrangement within a single structure (1EGH) that yielded pairwise similarities over 9.9. The citrate examples fell into three categories: surface-bound pairs of ligands bound to serine proteases (the bulk of cases), a trio of symmetrically bound citrate ligands bound to a macrophage inhibitory factor protein, and a pair bound to unrelated sites of a ribonuclease. Despite the heterogeneity of the last set, all segregated into a single subtree. The clustering tree reported by Yeturu et al.⁸ showed similar results, but grouped the trio of 1GCZ citrate sites within the phosphoglycolic acid subtree and apart from the remainder of the citrate binding sites.

Relationship Between Protein Pocket and Ligand Similarity

A natural expectation is that there should exist some degree of concordance between the similarity of protein pockets and the similarities exhibited by the ligands that bind them. There are two quite different cases within the data sets examined here. One is the case where many synthetic ligands have been designed to competitively bind a single protein's active site (e.g. the 29 different ligands of p38 α from Figure 4). The other is the case where nature has evolved multiple strategies for binding the same naturally occurring cognate ligands, which is the situation in the Kahraman Set. Figure 13 shows both relationships, with ligand similarity computed using the Surflex-Sim approach.²

The plot at left in Figure 13 shows ligand similarity versus protein pocket similarity for all pairs of p38 α comparisons. The relationship was statistically significant ($p \ll 0.01$ by Kendall's Tau, estimated by permutation analysis), though clearly stronger at the extremes of ligand similarity and dissimilarity. Recalling Figure 4, pairs of ligands such as those in 1WBT and 1WBS had very high similarity (9.4) with correspondingly high pocket similarity (9.3), whereas 1WBT compared with 1OVE yielded much lower ligand and protein similarity (4.9 and 6.9, respectively).

The plot at right in Figure 13 shows ligand similarity versus protein pocket similarity for the Kahraman data set, where the protein similarity values resulted from average all pairs of comparisons for the cognate proteins of each ligand type and the ligand similarity values were again computed using Surflex-Sim. With the exception of the glucose/phosphate binding site comparison, the relationship between ligand and protein similarity was nearly linear. Overall, the correlation was 0.46 by Kendall's Tau ($p \ll 0.01$, by permutation). Further, the relationship between PSIM values and ligand similarity was much stronger than

between PSIM values and volume similarities (Kendal's Tau of 0.30). Even when restricting the set of comparisons to protein pairs where volume differences no longer correlated with protein pocket similarities, the statistically significant relationship between pocket and ligand similarity remained.

The relationship between ligand similarity and pocket configuration within the p38 α variants was subtle. To test whether the PSIM approach was unique in its ability to identify this effect, we assessed whether the PocketMatch⁸ and SOIPPA¹⁸ methods yielded correlations with ligand similarity for the p38 α subset. Figure 14 shows the plots of ligand and protein similarity for the two methods. Neither algorithm was able to yield correlations between local pocket similarity and the similarities of bound ligands. When restricted to ligand similarities less than 7.0, PSIM yielded a statistically significant correlation (Tau = 0.10, $p < 0.01$), but both PocketMatch and SOIPPA yielded slightly negative correlations (the former with $p < 0.05$).

The two cases explored here do not fully examine the issues around the relationship between protein pocket similarity and ligand similarity. The p38 α set considered a relatively flexible protein pocket bound to a series of competitive ligands with a range of underlying scaffolding. Clearly, in the case of a very rigid protein, it would be more difficult to discern conformational effects and relate them to ligand structural patterns. There are many examples of proteins that undergo little change on binding structurally divergent ligands. The Kahraman Set was very different in character, with examples of multiple diverse proteins each bound to the same ligand. In this case, the degree of concordance between average pocket similarity and ligand similarity was striking, though the issue of binding site volume differences limits the generality of the observation. It bears mention as well that different approaches to the computation of ligand similarity would also affect the results. Similarity approaches that measure topological structural similarity between ligands, for example, might yield no relationship between pocket and ligand similarity. As with proteins sharing little sequence similarity but exhibiting similar binding pockets, small molecules may share very little topological similarity but exhibit very similar surface shape and polarity.

Relationship to Other Approaches

There are a number of protein structural comparison algorithms, broadly characterized by whether they are global or local, backbone-based or include sidechain information, and the degree to which they make comparisons based purely on shape or also based on polarity. The Surflex-PSIM approach is local, accounts for all protein atoms, and considers the detailed comparison of both shape and protein surface polarity. Other methodologies have approached protein similarity in a variety of ways such as geometric hashing, shape alignment, and fingerprinting. These various methods can be local or global and some produce alignments while others do not.

Direct comparisons were made here with five notable recent examples of pocket comparison algorithms: the SiteBase algorithm of Kinnings et al.,¹⁰ the spherical harmonics approach of Kahraman et al.,⁶ the sup-CK method of Hoffman et al.,⁷ the PocketMatch method of Yeturu et al.,⁸ the SOIPPA method of Xie et al.¹⁸ Each of these algorithms can compute local similarities, and each are capable of producing alignments between protein binding sites. In each case, comparisons were made either by analyzing the performance of PSIM on sets used by the authors of other methods (the Kinnings, Kahraman, Hoffman, and Yeturu sets), or by analyzing the performance of other methods on a set introduced in this study (PocketMatch and SOIPPA on the p38 α set). In the former comparisons, PSIM performed as well as the best reported methods, within the ability of statistics to distinguish among them.

In the latter comparison, while limited in scope, PSIM pocket similarities exhibited a unique relationship to bound ligand similarities.

Many more approaches for protein structure comparison have been developed, notably the Dali work of the Sander group^{19,20} and the combinatorial extension method from the Bourne group^{21–23}. These methods have been developed primarily for the study and characterization of the space of protein structures and the relationship of global structure to function, which is a different focus than the proposed work, which seeks to address local comparisons between protein surfaces. Closer to this concept are methods for 3D protein motif identification (e.g. SeqFEATURE²⁴ and GASPS.²⁵ These approaches identify local structural features within proteins (e.g. the catalytic triad of serine proteases) that establish a functional motif for proteins with related function. As with the global alignment methods, these have been developed primarily for characterization and annotation of protein function, but not to distinguish, for example, the differences in ligand specificity between two different serine proteases. A very different family of approaches also geared toward protein function annotation describes pockets based on fingerprint vectors, which do not yield alignments. A recent example called FuzCav was reported by Weill and Rognan.²⁶ The approach constructs a cavity descriptor to characterize a protein binding site based upon the counts of 4833 different pharmacophoric features. The approach is very fast, requiring only comparison of pre-computed vectors, and it was effective in distinguishing different classes of proteins. Our approach is geared specifically toward detailed pocket surface comparison based on joint alignment, and we view it as being complementary.

Because of PSIM's focus on surface shape and charge it can isolate small changes between protein structures that might go unnoticed with other methodologies. It is also the only method that shares its underlying formalism with a mature method for computing small molecule molecular similarity.^{1,2}

Conclusions

We have shown that a local, surface-based, protein pocket similarity metric yields informative results across several levels of protein inter-relatedness. Among closely related kinases (including many alternate conformations of single proteins), we showed that the PSIM metric grouped proteins bound to similar ligands more closely than those bound to more divergent small molecules. Among a more diverse set of kinases, we showed that kinase ligand binding specificity was related to our direct computation of protein pocket similarity, with proteins binding the same ligands having more similar pockets to one another (even in spite of quite different primary sequences) than proteins not sharing ligand binding preferences. Among an extremely diverse set of proteins, all of which bind ATP, we showed that the means by which ATP is bound varies and that different structural strategies can be identified purely based on local surface similarity. The structural motifs were strong enough that methods making use of even multiple sequences were unable to correctly identify the motifs. Among heterogeneous sets of proteins, where protein classes were represented by diverse exemplars (the Kahraman and Hoffman Sets) or by highly similar exemplars (the Yeturu Set), the PSIM approach performed at least as well as the best reported methods. Unique to PSIM was the correspondence between protein pocket similarity and ligand similarity.

Within each of these levels of protein comparison comes the opportunity for future application of the methodology. At the level of highly related proteins (e.g. conformational variants of a particular kinase), automated alignment and selection of conformational variants for molecular docking studies is of interest. Current approaches generally rely on single protein conformations for screening libraries of ligands, but addressing protein

conformational variability has clear benefits.²⁷ At the level of related families of proteins that are interesting from a drug discovery point of view (e.g. serine proteases or human kinases), careful comparison of active sites may help identify potential sources for off-target effects of small molecule therapeutics. Conceivably, non-obvious off-target effects could also be identified, given that sequence relatedness is not required for the method to identify strong structural motifs.

Among the broadest set of proteins, one of the most interesting possibilities lies in functional annotation. Here, there are two clear opportunities. The first combines the structural alignment approach with local sequence motif identification in the hope that the former amplifies the signal for the latter, enabling identification of as yet unknown sequence motifs that could be used to annotate functions for proteins whose structure has yet to be elucidated. The second would seek to comprehensively profile proteins whose structure has been determined specifically to help yield convincing functional information.²⁸

There are also significant technical areas in which further study will be required. Understanding the thresholds at which different levels of similarity support some level of confidence in making a functional annotation will require broader study of larger sets of proteins. It is likely that raw similarity values will be context-dependent in the sense that a particular similarity value computed against a large, complex, binding site would probably mean more than the same value computed against a smaller and less complex site. The method is also computationally expensive in its current implementation, requiring on the order of one minute per comparison on standard desktop hardware. The speed issue derives from the adaptation of this method from small molecules, where alignment optimization involves moving one ligand onto another. Applied in the most straightforward manner, this is inefficient for proteins owing to their large number of atoms. An adaptation of the approach where molecular observer points are moved, with proteins remaining fixed until a final alignment is produced, will yield a substantial speed benefit. Gains in computational throughput will support broader characterization of the PSIM approach and will offer more practical application of the method for prospective studies.

Acknowledgments

The authors gratefully acknowledge NIH for partial funding of the work (grant GM070481).

References

1. Jain AN. Morphological similarity: a 3D molecular similarity method correlated with protein-ligand recognition. *J Comput Aided Mol Des.* 2000; 14(2):199–213. [PubMed: 10721506]
2. Jain AN. Ligand-based structural hypotheses for virtual screening. *J Med Chem.* 2004; 47(4):947–961. [PubMed: 14761196]
3. Jain AN, Harris NL, Park JY. Quantitative binding site model generation: compass applied to multiple chemotypes targeting the 5-HT_{1A} receptor. *J Med Chem.* 1995; 38(8):1295–1308. [PubMed: 7731016]
4. Langham JJ, Cleves AE, Spitzer R, Kirshner D, Jain AN. Physical binding pocket induction for affinity prediction. *J Med Chem.* 2009; 52(19):6107–6125. [PubMed: 19754201]
5. Nicholls A, McGaughey GB, Sheridan RP, Good AC, Warren G, Mathieu M, Muchmore SW, Brown SP, Grant JA, Haigh JA, Nevins N, Jain AN, Kelley B. Molecular shape and medicinal chemistry: a perspective. *J Med Chem.* 2010; 53(10):3862–3886. [PubMed: 20158188]
6. Kahraman A, Morris RJ, Laskowski RA, Thornton JM. Shape variation in protein binding pockets and their ligands. *J Mol Biol.* 2007; 368(1):283–301. [PubMed: 17337005]
7. Hoffmann B, Zaslavskiy M, Vert JP, Stoven V. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3D: application to ligand prediction. *BMC Bioinformatics.* 2010; 11:99. [PubMed: 20175916]

8. Yeturu K, Chandra N. PocketMatch: a new algorithm to compare binding sites in protein structures. *BMC Bioinformatics*. 2008; 9:543. [PubMed: 19091072]
9. Fabian MA, Biggs WH 3rd, Treiber DK, Atteridge CE, Azimioara MD, Benedetti MG, Carter TA, Ciceri P, Edeen PT, Floyd M, Ford JM, Galvin M, Gerlach JL, Grotzfeld RM, Herrgard S, Insko DE, Insko MA, Lai AG, Lelias JM, Mehta SA, Milanov ZV, Velasco AM, Wodicka LM, Patel HK, Zarrinkar PP, Lockhart DJ. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat Biotechnol*. 2005; 23(3):329–336. [PubMed: 15711537]
10. Kinnings SL, Jackson RM. Binding site similarity analysis for the functional classification of the protein kinase family. *J Chem Inf Model*. 2009; 49(2):318–329. [PubMed: 19434833]
11. Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. Binding MOAD (Mother Of All Databases). *Proteins*. 2005; 60(3):333–340. [PubMed: 15971202]
12. Rose PW, Beran B, Bi C, Bluhm WF, Dimitropoulos D, Goodsell DS, Prlic A, Quesada M, Quinn GB, Westbrook JD, Young J, Yukich B, Zardecki C, Berman HM, Bourne PE. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic Acids Res*. 2011; 39(Database issue):D392–D401. [PubMed: 21036868]
13. Caetano-Anolles G, Kim HS, Mittenthal JE. The origin of modern metabolic networks inferred from phylogenomic analysis of protein architecture. *Proc Natl Acad Sci U S A*. 2007; 104(22):9358–9363. [PubMed: 17517598]
14. Hanks SK, Quinn AM, Hunter T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*. 1988; 241(4861):42–52. [PubMed: 3291115]
15. Meng EC, Pettersen EF, Couch GS, Huang CC, Ferrin TE. Tools for integrated sequence-structure analysis with UCSF Chimera. *BMC Bioinformatics*. 2006; 7:339. [PubMed: 16836757]
16. Hanks SK, Hunter T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. *FASEB J*. 1995; 9(8):576–596. [PubMed: 7768349]
17. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics*. 2007; 23(21):2947–2948. [PubMed: 17846036]
18. Xie L, Bourne PE. A unified statistical model to support local sequence order independent similarity searching for ligand-binding sites and its application to genome-based drug discovery. *Bioinformatics*. 2009; 25(12):i305–i312. [PubMed: 19478004]
19. Holm L, Sander C. Dali: a network tool for protein structure comparison. *Trends Biochem Sci*. 1995; 20(11):478–480. [PubMed: 8578593]
20. Holm L, Sander C. Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res*. 1997; 25(1):231–234. [PubMed: 9016542]
21. Shindyalov IN, Bourne PE. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*. 1998; 11(9):739–747. [PubMed: 9796821]
22. Shindyalov IN, Bourne PE. An alternative view of protein fold space. *Proteins*. 2000; 38(3):247–260. [PubMed: 10713986]
23. Shindyalov IN, Bourne PE. A database and tools for 3-D protein structure comparison and alignment using the Combinatorial Extension (CE) algorithm. *Nucleic Acids Res*. 2001; 29(1):228–229. [PubMed: 11125099]
24. Wu S, Liang MP, Altman RB. The SeqFEATURE library of 3D functional site models: comparison to existing methods and applications to protein function annotation. *Genome Biol*. 2008; 9(1):R8. [PubMed: 18197987]
25. Polacco BJ, Babbitt PC. Automated discovery of 3D motifs for protein function annotation. *Bioinformatics*. 2006; 22(6):723–730. [PubMed: 16410325]
26. Weill N, Rognan D. Alignment-free ultra-high-throughput comparison of druggable protein-ligand binding sites. *J Chem Inf Model*. 2010; 50(1):123–135. [PubMed: 20058856]
27. Jain AN. Effects of protein conformation in docking: improved pose prediction through protein pocket adaptation. *J Comput Aided Mol Des*. 2009; 23(6):355–374. [PubMed: 19340588]
28. Hendrickson WA. Impact of structures from the protein structure initiative. *Structure*. 2007; 15(12):1528–1529. [PubMed: 18073103]

Figure 1.

Figure 2.

Figure 3.

Figure 4.

Figure 5.

Full Clustering

GKST Loop Proteins

Type II TRNA Synthetases

Kinases

Figure 6.

Automated Alignment of Kinases

Figure 7.

Figure 8.

Structure Based Alignment of Kinases

Figure 9.

Automated Alignment of GKST Loop Bearing Proteins

Figure 10.

Sequence Alignments of ATP Cluster

PDB ID	Compound	Source	3D Based Alignment	ClustalW Alignment
1BoU	HISP	S. Typhimurium	G S S G S G K S T F L R C I	G S G K S T F L R C
1FMW	MHCA	D. Discoideum	G E S G A G K T E N T K K V	G A G K T E N T K K
1IIo	ARSA	E. Coli	G K G G V G K T - - - - -	S G A C T T E I A A
1KO5	GNTK	E. Coli	G V S G S G K S A V A S E V	G S G K S A V A S E
1L2T	MJo796	M. Jannaschii	G P S G S G K S T M L N I I	G S G K S T M L N I
1MV5	LMRA	L. Lactis	G P S G G G K S T I F S L L	G G G K S T I F S L
1NSF	NSF	C. Griseus	- P P H S G K T A L A A K I	- S D R T P L V S V
1Q12	MALK	E. Coli K12	G P S G C G K S T L L R M I	G C G K S T L L R M
1QHX	CPT	S. Venezuelae	G G S S A G K S G I V R C L	S A G K S G I V R C
1RoX	CFTR	M. Musculus	G S T G S G K T S L L M L I	G S G K T S L L M L
1VCI	PHo022	P. Horikoshii	G P S G C G K T T T L R M I	G C G K T T T L R M
1XEF	HLYB	E. Coli	G R S G S G K S T L T K L I	G S G K S T L T K L
1XEX	SMC	P. Furiosus	G A N G S G K S N I G D A -	G S G K S N I G D A
1XKV	PCKA	T. Thermophilus	G L S G T G K T T L S T - -	P F A P E A F E A L
1YTM	PCK1	A. Succiniciproducens	G L S G T G K T T L S T - -	P V T E E A W A Q L
2A5Y	CED9	C. Elegans	G R A G S G K S V I A S Q A	Q F S R Q M L D R K
2BBS	CFTR	H. Sapiens	G S T G A G K T S L L M M I	G A G K T S L L M M
2C8V	NIFH	A. Nelandii	G K G G I G K S - - - - -	P K A D S T R L I L
2CBZ	MRP1	H. Sapiens	G Q V G C G K S S L L S A L	G C G K S S L L S A
2EWW	PILT	A. Aeolicus	G P T G S G K S T T I A S -	A V R I D G Y I K F
2IXE	TAP2	R. Norvegicus	G P N G S G K S T V A A L L	G S G K S T V A A L
2IYW	AROK	M. Tuberculosis	G L P G S G K S T I - - - -	G S G K S T I G R R
2OLR	PCK1	E. Coli K12	G L S G T G K T T L S T - -	P L S P E T W Q H L
2PZE	CFTR	H. Sapiens	G S T G A G K T S L L M M I	G A G K T S L L M M
2R9V	ATP1	T. Maritima MSB8	G D R Q T G K T A I A I D T	P V G E E L L G R V
2V7Q	ATP1	B. Taurus	G D R Q T G K T S I A I D T	P V G E E L L G R V
3FKQ	NTRC-like	E. Rectale	P C G G V G T S T V A A A C	K Y A D K L E V Y S
3FVQ	FBPC	N. Gonorrhoeae	G A S G C G K T T L L R C L	G C G K T T L L R C

Figure 11.

PSIM: Ligand vs. Protein Similarity

Figure 13.

Alternate Methods: p38 α Ligand vs. Protein Similarity

Figure 14.

Comparison of PSIM to results from Kinnings et al. The table shows enrichment for structures whose corresponding enzyme was known to bind the cognate ligand of the query PDB structure. The results were nearly identical (minor differences in the numbers of proteins were due to technical differences in the protocols). Note, however, that enrichment values (e.g. for SB203580, BIRB-796, and Roscovitine) are dominated by multiple structural variants of the cognate enzyme being present in the analysis (see text for details). For Tarceva, by contrast, the highly ranked structures included some with low sequence similarity to EGFR (the protein in 1M17).

Table 1

Ligand	PDB Code	Actives in top 5%	Total Actives	Enrichment (max. possible)	Kinnings Enrichment	PSIM p-value
Tarceva	1M17	8	24	6.1 (7.6)	6.9 (9.2)	1.10E-06
Gleevec	1OPJ	5	17	5.4 (10.7)	6.0 (12.0)	7.40E-04
SB203580	1A9U	10	39	4.7 (4.7)	5.4 (5.4)	7.40E-08
BIRB-796	1KV2	10	53	3.4 (3.4)	3.7 (4.0)	2.30E-06
Gleevec	1T46	4	17	4.3 (10.7)	3.0 (12.0)	7.90E-03
Roscovitine	1UNG	9	80	2.1 (2.3)	2.8 (2.8)	3.00E-03
SP600125	1UKI	7	92	1.4 (2.0)	1.8 (2.4)	1.70E-01
SP600125	1PMV	7	92	1.4 (2.0)	1.6 (2.4)	1.70E-01
GW-2016	1XKK	1	1	18.2 (18.2)	0 (19.9)	5.50E-02