# Binding Site Similarity Analysis for the Functional Classification of the Protein Kinase Family

Sarah L. Kinnings and Richard M. Jackson*

Institute of Molecular and Cellular Biology and Astbury Centre for Structural Molecular Biology, Faculty of Biological Sciences, University of Leeds, Leeds LS2 9JT, U.K.

Received August 19, 2008

Methods for analyzing complete gene families are becoming of increasing importance to the drug discovery process, because similarities and differences within a family are often the key to understanding functional differences that can be exploited in drug design. We undertake a large-scale structural comparison of protein kinase ATP-binding sites using a geometric hashing method. Subsequently, we propose a relevant classification of the protein kinase family based on the structural similarity of its binding sites. Our classification is not only able to reveal the great diversity of different protein kinases and therefore their different potential for inhibitor selectivity but it is also able to distinguish subtle differences within binding site conformation reflecting the protein activation state. Furthermore, using experimental inhibition profiling, we demonstrate that our classification can be used to identify protein kinase binding sites that are known experimentally to bind the same drug, demonstrating that it has potential as an inverse (protein) virtual screening tool, by identifying which other sites have the potential to bind a given drug. In this way the cross-reactivities of the anticancer drugs Tarceva and Gleevec are rationalized.

## INTRODUCTION

Constituting 518 genes and around 1.7% of the human genome, the protein kinase family is one of the largest protein families.[1] All eukaryotic protein kinases are the result of divergent molecular evolution from a single ancestral protein, and they share a conserved catalytic domain of approximately 300 amino acids.[2] This domain is responsible for a common catalytic mechanism: the transfer of the $\gamma$-phosphate of adenosine triphosphate (ATP) to serine, threonine, or tyrosine residues on protein substrates.[3] The phosphorylation of proteins is a universal mechanism for the regulation of key cellular processes, including cell growth and differentiation, apoptosis, metabolic pathways, and membrane transport.[4] Indeed, phosphorylation can serve to activate, inhibit, or cause conformational changes in proteins, which in turn can stimulate a wide range of responses in the cell.[3]

The spatial and temporal control of phosphorylation depends on the correct regulation of protein kinases.[5] Each protein kinase responds to a unique set of intra- and extracellular signals that modulates its activity.[6] The integrity of signal transduction relies on the fact that each protein kinase is highly specific, acting only on a defined subset of cellular targets.[7] During signaling events only a small proportion of the protein kinase population is activated. These active protein kinases are rapidly inactivated by cellular regulatory mechanisms. Control of the inactive state is essential in order to ensure signaling fidelity. Indeed, the constitutive activation of protein kinases, which can be caused by the failure of regulatory mechanisms, mutations, or overexpression, is a key factor in the development of human diseases such as cancer, diabetes, and inflammatory disorders.[8]

The targeted inhibition of protein kinases represents a key therapeutic strategy for the treatment of a diverse range of pathological conditions. Therefore, the regulation of protein kinase activity is a principle focus of current pharmaceutical research.[8] In fact, it is estimated that around one-third of current drug discovery programs target protein kinases.[9] In a recent review, it was reported that a total of 113 mature chemical compounds have been developed as protein kinase inhibitors.[10] In the past decade, breakthrough advances have led to the approval of several small molecule kinase inhibitors for clinical use. The success of these drugs, which include imatinib mesylate ('Gleevec', Novartis, 2001), gefitinib ('Iressa', AstraZeneca, 2003), erlotinib ('Tarceva', Genentech and OSIP, 2004), sorafenib tosylate ('Nexavar', Bayer and Onyx, 2005), sunitinib malate ('Sutent', Pfizer, 2006), dasatinib ('Sprycel', Bristol-Myers Squibb, 2006), lapatinib ('Tyverb', GlaxoSmithKline, 2007), and nilotinib ('Tasigna', Novartis, 2007), has encouraged further investment in the field.[9]

The vast majority of protein kinase inhibitors that are currently available target the ATP-binding site. The evolutionary design features of the ATP-binding site are advantageous for the binding of ATP competitive inhibitors. Indeed, unlike the peptide binding site, it is well-structured and deep; engulfing nearly the entire ATP molecule (only the $\gamma$-phosphate is left exposed and available for the phosphotransfer reaction). However, its high degree of sequence and structural conservation among the protein kinases means that the design of selective kinase inhibitors is a significant challenge.[4,9] In fact, no ATP mimetics have been developed that inhibit only a single protein kinase.[11] However, this lack of selectivity

* Corresponding author phone: +44 (0)113 343 2592; fax: +44 (0)113 343 3167; e-mail: r.m.jackson@leeds.ac.uk.

10.1021/ci800289y CCC: $40.75 
Published on Web 01/13/2009

can sometimes be advantageous since anticancer drugs that act on multiple tyrosine kinases are thought to be more effective than those that are specific in their mode of action.[12] Cross-reactivity with unrelated ATP-binding proteins is also a frequent problem, resulting in undesirable side effects that cause many protein kinase inhibitors to fail in either preclinical or clinical development.[2]

Considerable efforts have been made to classify the protein kinase family based on various different types of information with a view to gaining a better insight into selectivity filters that can be used to discriminate between different protein kinases. In 2002, Manning and co-workers provided an extensive annotation of the protein kinase family when they classified 518 nonredundant protein kinase genes in the human genome based on their sequence similarity.[13] In the same year, Cheek et al. combined both sequence and structural information for their classification of the kinase superfamily.[14,15] In 2004, Vieth et al. presented a classification of protein kinases based on small-molecule inhibition data which they extracted from the literature.[16] Recently, experimental approaches have been used to produce interaction profiles for protein kinase inhibitors.[17−19] Such experimental profiling facilitates the identification of unknown inhibitor targets as well as predicting cross-reactivity with other protein kinases.

Other classification approaches have focused on the three-dimensional structure or the physiochemical properties of protein kinase binding sites. Indeed, the rapid expansion of structural genomics efforts in recent years has allowed the field of protein kinase structural biology to develop quickly. During 2005 and 2006 alone, the crystal structures of more than half of all novel protein kinase catalytic domains were solved. To date, over 100 unique crystal structures of protein kinase catalytic domains have been published in the Protein Data Bank (PDB).[20] In 2002, Naumann and Matter used this structural information to classify a set of 26 different protein kinases into subfamilies based on the similarity of their protein−ligand interactions. After superimposing the kinase binding sites, they used various methods to identify features that were common to the different binding sites as well as those regions that differed among the subfamilies. In this way, the authors were able to carry out a classification of the protein kinases and identify the structural features important to each subfamily.[21] Last year, Kuhn et al. proposed a classification of a diverse set of 258 protein kinase cavities in terms of functional recognition patterns in their active sites. Their classification enabled them to detect cross-relations among protein kinases unrelated in sequence, which they then rationalized using small-molecule inhibition data.[22]

Here we report the comparison of a nonredundant set of 354 protein kinase binding sites using a geometric hashing algorithm.[23] The geometric hashing algorithm compares binding sites on the basis of their atom-atom similarity alone and does not rely on any other characteristics such as shape, amino acid composition, or sequence similarity information. Subsequently, a classification of these binding sites is presented and compared with one that is sequence-based. Unexpected binding site similarities and dissimilarities are identified and discussed. In addition, a more detailed analysis of the classification of the binding sites of mitogen-activated protein (MAP) kinase p38α, cyclin-dependent kinase 2 (CDK2), and Abelsone tyrosine kinase (c-Abl) is performed. Finally, binding site similarity is compared with data from experimental inhibition profiling to determine whether it is able to capture similarities between sites that can bind the same compound. This is found to be the case for almost all of the known inhibitors for which structural data and inhibition profiling is known. In this way the cross-reactivity of the antileukemia drug Gleevec is rationalized.

## METHODS

**Compilation of Data.** In order to compile a data set of protein kinase crystal structures, PDB codes were retrieved from the Pkinase (PF00069) and Pkinase_Tyr (PF07714) families of the Pfam protein families database.[24] 395 PDB codes were retrieved from the Pkinase family, and 81 were retrieved from the Pkinase_Tyr family. To eliminate redundancy from the data set, only crystal structures with unique protein−ligand pairs were retained. Similarly, only one apo form of each protein was retained. Where there were multiple structures of the same protein bound with the same ligand, or of the same protein in apo form, the structure with the highest resolution was chosen for the data set. Each ligand was investigated to ensure that it was biologically active and bound in the ATP-binding site. The additional removal of theoretical models resulted in a data set consisting of 354 protein kinase crystal structures, constituting a total of 78 different proteins (or 75 if the source organism is not taken into account) (see the Supporting Information, Table SI 1). Many of the proteins were represented multiple times in the data set, with several different ligands bound. A total of 301 of the structures had been crystallized with an inhibitor, and the remaining 53 had been crystallized in apo form.

**Extraction of Binding Sites.** A single protein chain corresponding to the selected ligand was extracted from each of the 354 structures (or if there was no ligand, the first protein chain in the PDB file was chosen). A geometric hashing algorithm[23] was used to superimpose all of the protein chains onto that of a cAMP-dependent protein kinase (PDB code: 1atp), a typical protein kinase domain, based on C-alpha backbone atoms only. Although the ligands were superimposed along with the structures, the geometric hashing algorithm ignored the ligand atoms during the superimposition process. A compound ligand consisting of all of the ligand atoms from the aligned structures was then created. Atoms were removed from the compound ligand if less than fifty other ligands (out of a possible 300) had atoms within a 1.5 Å radius. This threshold was introduced to remove sparsely populated regions of the compound ligand. Binding sites were then defined as all protein atoms within a 5 Å radius of any of the atoms in the compound ligand (Figure 1). All 354 binding sites were extracted in this way.

**Comparison of Binding Sites.** The geometric hashing method[23] can be used to rapidly compare a set of binding sites in an all-against-all manner. The algorithm proceeds by identifying equivalent heavy atom constellations between pairs of binding sites. Matching atom-atom correspondences occur in the same relative spatial orientation and have the same element (e.g., all carbon atoms are treated as equivalent). The atom-atom score is then defined as the number of atoms comprising the largest possible matching constellation ($n_{match}$). Subsequently, a similarity score, $R_3$, is calculated analogously to the Tanimoto coefficient.[25] The $R_3$ score takes into account the total size of the two binding sites ($n_{site1}$ and

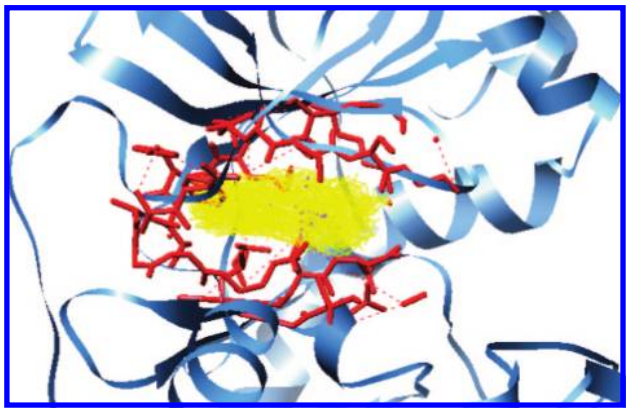

**Figure 1.** Definition of a binding site in cAMP-dependent protein kinase (PDB code: 1atp). Atoms of the compound ligand are shown in yellow, and the defined binding site is shown in red stick representation. The remainder of the protein is depicted as a blue ribbon.

$n_{site2}$) (see equation below) and can take any value in the range zero to one, where a value of one signifies the comparison of a binding site with itself.

$$R_3 = \frac{n_{match}}{n_{site1} + n_{site2} - n_{match}}$$

The geometric hashing algorithm was used to compare the 354 protein kinase binding sites in an all-against-all manner. An all-atom representation of the binding sites was used, so that no information regarding possible hydrogen bonds or electrostatic or steric interactions with the ligand would be lost (note that ligand coordinates were not included in the matching).

**Structure-Based Binding Site Clustering.** Binding site similarity scores were clustered using the *scluster* program in the clustering toolkit CLUTO.[26] Optimal clustering parameters were determined using a binding site clustering validation process benchmarked using 12 different enzyme (E.C.) families (see the Supporting Information). As a result, the agglomerative algorithm was chosen to cluster the $R_3$ scores into 78 different output clusters. The agglomerative algorithm uses the Unweighted Pair Group Method with Arithmetic mean (UPGMA) merging criterion. It initially considers every object as an individual cluster and sequentially merges the two most similar clusters until the number of required output clusters is obtained. 78 output clusters were chosen because there were 78 different proteins in the data set.

**Sequence-Based Clustering.** The sequences of the 354 protein kinases in the data set were retrieved from the PDB in FASTA format. In each case, only the sequence of the protein chain from which the binding site was extracted was retained. The part of the sequence corresponding to the protein kinase domain, as defined by the Universal Protein Resource (UniProt),[27] was then extracted. The ClustalW[28] multiple alignment algorithm was used to align all of the sequences in a pairwise manner, and the resulting sequence identities were clustered using the agglomerative algorithm and 78 output clusters.

**Unexpected Binding Site Similarities and Dissimilarities.** The mean sequence identity of all protein kinases in the data set was calculated to be 30.1%, and the mean $R_3$ score was calculated to be 0.25. In order to identify protein kinases that exhibited unexpected binding site similarity, pairs of protein kinases with an $R_3$ score in the top 5% of all $R_3$ scores ($R_3 > 0.449$) and a sequence identity of less than or equal to 30% were listed. Pairs of protein kinases were only retained if they appeared as pairs in the list at least three times, and the binding site pair exhibiting the highest $R_3$ score was chosen as the representative for that particular protein kinase pair. In this way, binding site pairs that were only sparsely populated in the list were excluded. In order to identify protein kinases that exhibited unexpected binding site dissimilarity, the same procedure was used, except that pairs of protein kinases with an $R_3$ score in the bottom 5% of all $R_3$ scores ($R_3 < 0.126$) and a sequence identity of more than 30% were listed.

**Enrichment Factors.** Enrichment factors were calculated in order to compare our binding site-based classification with 10 $\mu$M screening data from both Fabian et al.[18] and Invitrogen.[29] For each comparison, all binding sites of protein kinases that were not included in the *in vitro* screening panel were removed from our data set. Subsequently, binding sites in our data set that had been cocrystallized with any of the screened inhibitors were selected as 'query' binding sites. For each query binding site, all other binding sites in our data set were ranked according to their $R_3$ score with that particular query site. Enrichment factors were then calculated from this ranked data set as follows

$$Enrichment\ factor = \frac{Ah/Th}{(A/T)}$$

where $Ah$ = no. of binding sites of actives in the top 5% of the data set, $Th$ = total no. of binding sites in the top 5% of the data set, $A$ = total no. of binding sites of actives in the data set, and $T$ = total no. of binding sites in the data set.

P-values were determined from the hypergeometric distribution for all enrichment factors greater than 1.0.

**Consensus Scoring.** Consensus scoring was also used to analyze the similarity between binding sites. For every binding site in the data set, the mean similarity score with all binding sites of a particular protein kinase was calculated. Subsequently, the entire data set was reranked based on consensus scores.

## RESULTS

**Comparison of Sequence-Based and Binding Site-Based Classifications.** Figures 2 and 3 show the sequence-based classification colored according to sequence similarity and according to binding site-based similarity, respectively. The degree of red coloration indicates the level of similarity according to each attribute. The two plots reveal a similar overall pattern, suggesting that protein kinases that are highly related in sequence also have structural site similarity. One of the most striking differences between the two plots is the dissimilarity of the MAP kinase p38α (p38α) binding sites to most other protein kinase binding sites. In particular, the p38α binding sites show a strong dissimilarity to those of the other MAP kinases and the CDK2s, to which they are closely related in terms of sequence. Such dissimilarities between protein kinase binding sites might be exploited in the development of selective protein kinase inhibitors. Indeed, numerous highly selective p38α inhibitors have been reported.[30−32] Conversely, the cell-cycle checkpoint kinase

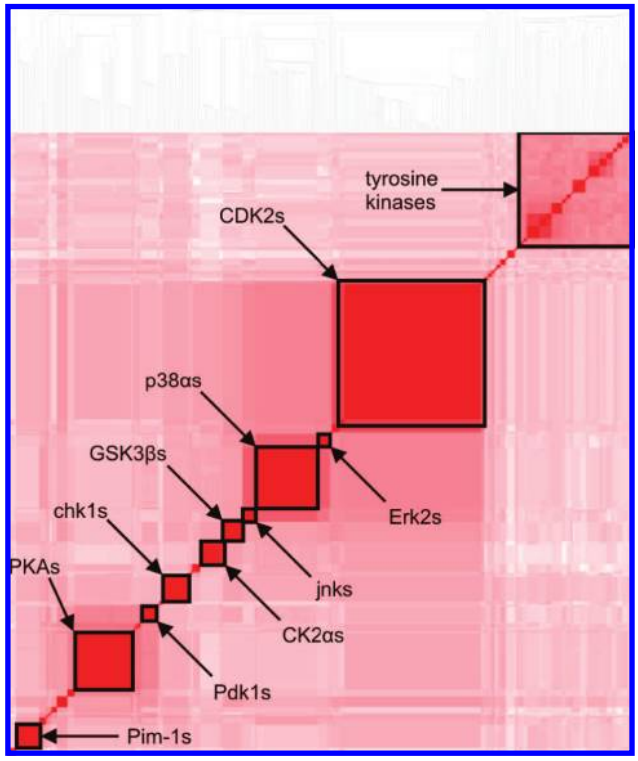


**Figure 2.** Sequence-based classification of the protein kinase family. The level of sequence similarity is indicated by the intensity of the red coloring. Major groupings have been labeled (see the Supporting Information for full clustering solution).

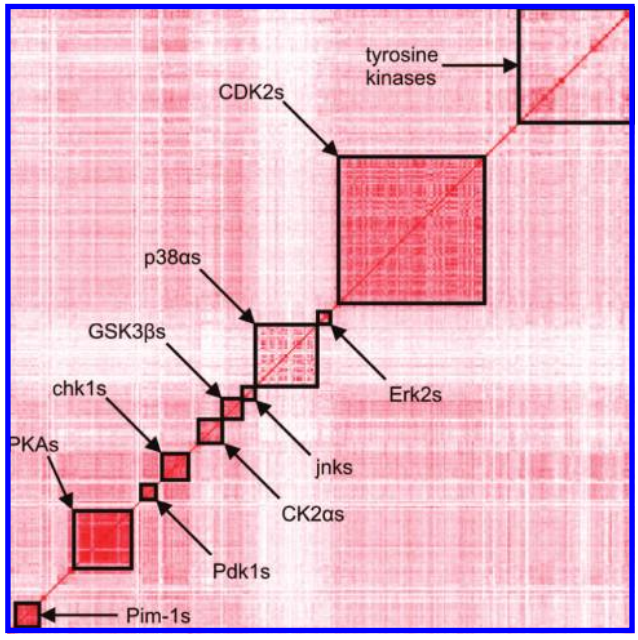


**Figure 3.** Sequence-based classification colored according to binding-site similarity. Similarity is indicated by the intensity of the red coloring. Major groupings are labeled in the same way as in Figure 2.

1s (chk1s) appear to be more similar to many other protein kinases in terms of their binding sites than their sequence similarity would suggest. Such unexpected similarities between binding sites may have implications for inhibitor cross-reactivity.

78 output clusters were chosen for the sequence-based classification in Figure 2 because 78 proteins were represented in the data set (lines discriminating between the

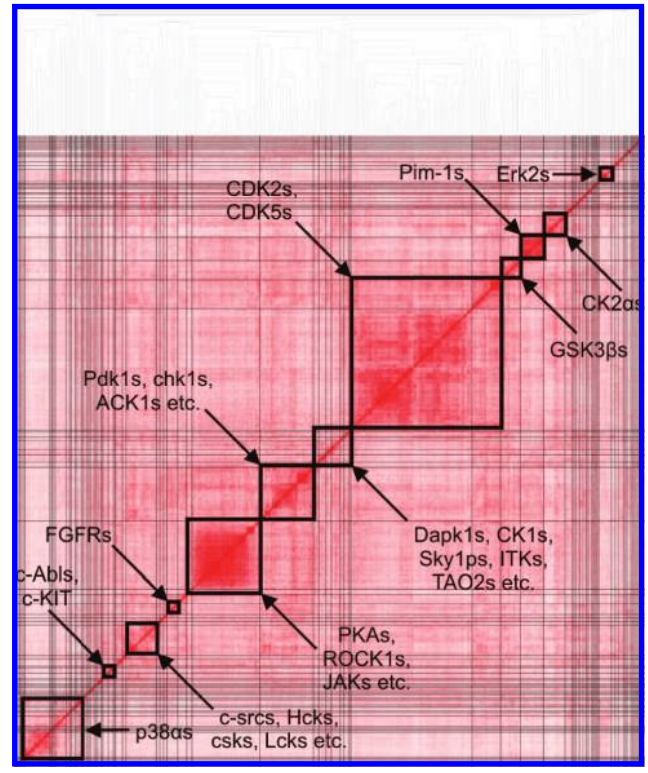


**Figure 4.** Structure-based binding site classification of the protein kinase family. Binding site similarity is indicated by the intensity of the red coloring. Major groupings have been labeled (see the Supporting Information for full clustering solution).

clusters have been removed for clarity). In this classification, the kinases cluster precisely according to their protein types. Figure 4 shows the structure-based binding site classification of the protein kinases. This classification appears much less organized, as might be expected since the 78 output clusters have little meaning, and protein kinases are in many cases mixed between different clusters.

In the structure-based binding site classification, there are nine different protein kinases for which a single cluster has been generated containing all binding sites of that kinase and no binding sites from other kinases. These are the proto-oncogene serine/threonine-protein kinase (Pim-1), glycogen synthase kinase-3 beta (Gsk3b), MAP kinase Erk2 (Erk2), death-associated protein kinase (Dapk1), hepatocyte growth factor receptor (c-MET), interferon-induced dsRNA-activated protein kinase, MAP kinase 8 (jnk1), mycobacterial protein kinase PknB, and tyrosine protein kinase Itk/Tsk. There are also several cases where related protein kinases have been clustered together. For instance, the epha2 and ephb2 receptor tyrosine kinases form their own cluster, as do the dual specificity MAP kinases Mek1 and Mek2, and the fibroblast growth factor receptor 1 and 2 (FGFR1 and 2) protein kinases. This demonstrates that the structural similarity is sufficient, even in this highly conserved set of ATP binding sites, to achieve a reasonably accurate protein kinase classification using the geometric hashing method.

There are also fourteen different protein kinases whose members are all consecutive in the clustering order but either share their cluster with other protein kinases or reside in more than one cluster. For instance, the 3-phosphoinositide dependent protein kinase 1 (Pdk1) binding sites share a cluster with those of the chk1s and the activated CDC42 kinase 1s

**Table 1.** Unexpected Similarities in the Binding Sites of Protein Kinases[a]

| protein | Pfam family | protein | Pfam family | $R_3$ score | sequence identity (%) |
|---|---|---|---|---|---|
| **cell cycle checkpoint kinase chk1** | Pkinase | tyrosine-protein kinase ZAP-70 | Pkinase_Tyr | 0.640 | 20 |
| **cell cycle checkpoint kinase chk1** | Pkinase | Wee1-like protein kinase | Pkinase | 0.632 | 19 |
| **cell cycle checkpoint kinase chk1** | Pkinase | 3-phosphoinositide dependent protein kinase-1 Pdk1 | Pkinase | 0.616 | 24 |
| **cell cycle checkpoint kinase chk1** | Pkinase | aurora-related kinase 1 (aurora-2) | Pkinase | 0.594 | 29 |
| **cell cycle checkpoint kinase chk1** | Pkinase | lymphocyte kinase (lck) | Pkinase_Tyr | 0.587 | 20 |
| cyclin-dependent PK, CDK2 | Pkinase | protein kinase C, theta type | Pkinase | 0.561 | 28 |
| **cell cycle checkpoint kinase chk1** | Pkinase | glycogen phosphorylase kinase (Phk), gamma-subunit | Pkinase | 0.559 | 26 |
| 3-phosphoinositide dependent protein kinase-1 Pdk1 | Pkinase | tyrosine-protein kinase SYK | Pkinase_Tyr | 0.558 | 13 |
| **cell cycle checkpoint kinase chk1** | Pkinase | cAMP-dependent PK, catalytic subunit | Pkinase | 0.546 | 26 |
| **cell cycle checkpoint kinase chk1** | Pkinase | cyclin-dependent PK, CDK2 | Pkinase | 0.538 | 28 |
| cAMP-dependent PK, catalytic subunit | Pkinase | tyrosine-protein kinase JAK3 | Pkinase_Tyr | 0.520 | 20 |
| **cell cycle checkpoint kinase chk1** | Pkinase | tyrosine-protein kinase SYK | Pkinase_Tyr | 0.516 | 18 |
| death-associated protein kinase, Dapk1 | Pkinase | 3-phosphoinositide dependent protein kinase-1 Pdk1 | Pkinase | 0.503 | 26 |
| protein kinase CK2, alpha subunit | Pkinase | cyclin-dependent PK, CDK2 | Pkinase | 0.502 | 30 |
| **cell cycle checkpoint kinase chk1** | Pkinase | fibroblast growth factor receptor 1 | Pkinase_Tyr | 0.500 | 19 |

[a] Pairs of protein kinases with $R_3$ scores in the top 5% (of all $R_3$ scores) and sequence identities of less than or equal to 30% (the mean sequence identity) are shown. The data have been ordered by descending $R_3$ score, and only the top fifteen pairs are shown. Pairs of kinases were only retained if their binding sites appeared as pairs in the list at least three times.

(ACK1s) in addition to a number of other protein kinases. Similarities between these protein kinases are much more subtle in the sequence-based classification. Conversely, the p38α binding sites have been split into twelve adjacent clusters, therefore indicating considerable dissimilarities among binding sites of this protein bound with different ligands.

**Unexpected Binding Site Similarities and Dissimilarities.** When high binding site similarity is accompanied by low sequence identity, there is the potential for unexpected ligand cross-reactivity. Therefore, knowledge of binding site similarities between protein kinases that are not closely related in sequence may aid in understanding the experimental results of the selectivity profiling for small molecule inhibition. Several protein kinase pairs exhibit a 'below average' sequence identity of 30% or less yet an $R_3$ score that is in the top 5% of all $R_3$ scores (Table 1). In some cases, binding site similarities are even observed between protein kinases belonging to different Pfam families. This suggests that the selectivity profiling of potential drug leads should not be restricted to kinases belonging to the same Pfam family.

It is interesting to note that the chk1s appear multiple times in the table (highlighted in bold). Chk1 is a key regulator of the cell cycle G2/M checkpoint, therefore making it an attractive target for the development of small molecule inhibitors against cancer. However, only a few chk1 inhibitors have been reported, most of which are natural products such as debromohymenialdisine (DBH) and staurosporine and its derivatives.[33] The lack of success in developing selective chk1 inhibitors may be partly explained by the fact that the chk1 binding sites exhibit such strong similarities to a large number of unrelated protein kinases, as detected by the geometric hashing method. Nevertheless, Foloppe et al. recently discovered that it is possible to design highly potent and selective chk1 inhibitors through the exploitation of a buried pocket at the periphery of the ATP-binding site.[33]

There are several protein kinase pairs in Table 1 for which there are structures in the data set that are bound with the same inhibitor. Although these pairs exhibit low sequence identity, similarities in their binding sites facilitate the binding of a common inhibitor, or more plausibly, the ligand induces a common conformational state in binding sites with sufficient plasticity i.e. an 'induced fit'. For instance, there are three structures in the data set that are cocrystallized with 7-hydroxystaurosporine; a chk1 structure, a Pdk1 structure, and a CDK2 structure. Interestingly, both chk1 and Pdk1 and chk1 and CDK2 are paired together in the table. Similarly, CDK2 and protein kinase CK2 alpha (CK2α), which also form a pair in the table, are both cocrystallized with the inhibitor 4,5,6,7-tetrabromobenzotriazole in the data set. Another example is that of the pairing between chk1 and Wee1-like protein kinase (Wee1). Like chk1, Wee1 is a cell cycle checkpoint kinase that controls entry into mitosis. It therefore has considerable potential as a drug target in the treatment of cancer.[34] Although these two kinases only have a sequence identity of 19%, Smaill et al. recently developed a series of potent inhibitors with dual specificity for Wee1 and chk1.[35]

When high sequence identity is accompanied by low binding site similarity, there is unexpected binding site dissimilarity that is not evident from sequence analysis alone. This may be important in highlighting where a protein exhibits a different binding site conformation, perhaps due to its activation state or as a result of ligand binding. Several protein kinase pairs exhibit an 'above average' sequence identity of greater than 30% yet an $R_3$ score that is in the bottom 5% of all $R_3$ scores (Table 2).

The p38α protein kinase appears multiple times in Table 2 (highlighted in bold). Not only does it show binding site dissimilarity to the closely related MAP kinase jnk1s, jnk3s, and Erk2s but also it shows dissimilarity to the CDK2/5/6/7s (which are in the same CMGC sequence group, as defined by Manning et al.[13]) and even dissimilarity to itself. In fact, a total of 26 pairs of different combinations of p38α binding

**Table 2.** Unexpected Dissimilarities in the Binding Sites of Protein Kinases[a]

| protein | Pfam family | protein | Pfam family | $R_3$ score | sequence identity (%) |
|---|---|---|---|---|---|
| **MAP kinase p38α** | Pkinase | cyclin-dependent PK, CDK5 | Pkinase | 0.074 | 36 |
| **MAP kinase p38α** | Pkinase | cyclin-dependent PK, CDK2 | Pkinase | 0.081 | 38 |
| **MAP kinase p38α** | Pkinase | mitogen-activated protein kinase 8 (jnk1) | Pkinase | 0.086 | 53 |
| **MAP kinase p38α** | Pkinase | **MAP kinase p38α** | Pkinase | 0.087 | 100 |
| **MAP kinase p38α** | Pkinase | MAP kinase Erk2 | Pkinase | 0.088 | 51 |
| **MAP kinase p38α** | Pkinase | c-jun N-terminal kinase (jnk3s) | Pkinase | 0.097 | 52 |
| cyclin-dependent PK, CDK2 | Pkinase | pak1 autoregulatory domain | Pkinase | 0.098 | 32 |
| Pkb kinase (Akt-2) | Pkinase | cAMP-dependent PK, catalytic subunit | Pkinase | 0.103 | 45 |
| abelsone tyrosine kinase (abl) | Pkinase_Tyr | insulin receptor | Pkinase_Tyr | 0.104 | 40 |
| **MAP kinase p38α** | Pkinase | cyclin-dependent PK, CDK6 | Pkinase | 0.105 | 35 |
| calmodulin-dependent protein kinase | Pkinase | 3-phosphoinositide dependent protein kinase-1 Pdk1 | Pkinase | 0.105 | 35 |
| **MAP kinase p38α** | Pkinase | cyclin-dependent PK, CDK7 | Pkinase | 0.116 | 36 |
| Pkb kinase (Akt-2) | Pkinase | aurora-related kinase 1 (aurora-2) | Pkinase | 0.120 | 33 |

[a] Pairs of protein kinases with $R_3$ scores in the bottom 5% (of all $R_3$ scores) and sequence identities of more than 30% (the mean sequence identity) are shown. The data have been ordered by ascending $R_3$ score. Pairs of kinases were only retained if their binding sites appeared as pairs in the list at least three times.

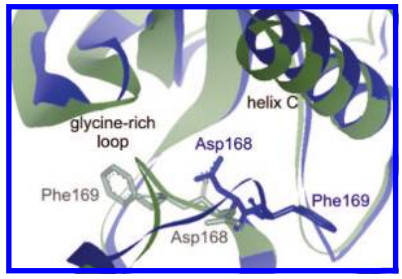


**Figure 5.** Comparison of the DFG-in and DFG-out conformations in two p38α structures. 1wbo (blue) is in the DFG-in conformation, whereas 1wbt (gray) is in the DFG-out conformation.

sites were found to have an $R_3$ score in the bottom 5% of all $R_3$ scores. The lowest $R_3$ score between two p38α binding sites was determined to be 0.087. Since these two kinase structures (PDB codes: 1a9u and 2baj) were both crystallized in the inactive state, this demonstrates the considerable variability in the inactive conformation of the p38α binding site.

**Binding Site-Based Classification: Individual Protein Kinases.** *MAP Kinase p38α.* The DFG (Asp-Phe-Gly) motif is a conserved feature of protein kinases which is able to undergo dramatic conformational changes as a protein kinase switches between inactive and active states (Figure 5). In the active state, the DFG motif adopts a so-called 'DFG-in' conformation, which is required for catalysis. The aspartate residue (Asp168) points toward the ATP-binding site, where it helps to orient the ATP $\gamma$-phosphate for phosphotransfer.[9] The phenylalanine residue (Phe169) makes contact with helix C, thereby assisting the formation of a Lys-Glu ion pair that is essential for catalysis. In inactive protein kinases, the DFG motif can exist in either a DFG-in or a DFG-out conformation. In the DFG-out conformation, the DFG main chain is shifted toward the ATP-binding site. The aromatic ring of its phenylalanine residue extends toward the ATP-binding site, whereas the aspartate side chain is positioned in the back cleft.[9]

Although all of the p38α binding sites in the data set were crystallized in the inactive conformation, they have been divided into twelve adjacent clusters in the binding site-based classification (Figure 6). This may be partly explained by the fact that, unlike many other protein kinases, the p38αs

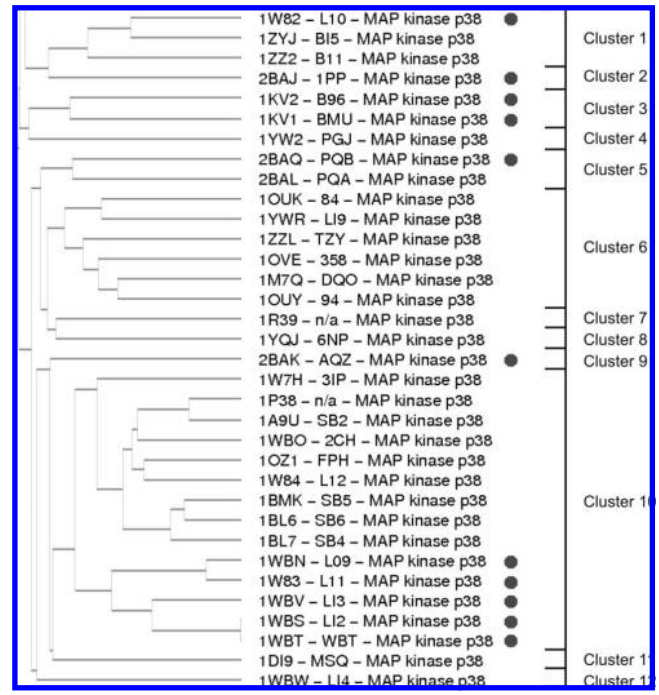


**Figure 6.** Close up view of the p38α kinases taken from the binding site-based classification shown in Figure 4. Binding sites that were crystallized in a DFG-out conformation have been marked with a gray dot.

are able to adopt inactive DFG-in and DFG-out conformations, thereby adding to their structural diversity. A total of eleven p38α structures in the data set were solved in the DFG-out conformation (these are marked with a gray dot in Figure 6). Some clustering of the DFG-out binding sites is apparent from the binding site-based classification, although this is further complicated due to the wide variety of conformational differences between the ligand bound states. In general, the p38α binding sites appear to cluster according to the type of inhibitor bound, suggesting that they are highly susceptible to ligand-induced conformational change. The following are a few examples:

- 1kv1 and 1kv2, which respectively bind the diaryl ureas BMU and BIRB 796, form their own cluster (cluster 3). Three structural changes were made to BIRB 796 in order to increase its binding affinity by 12,000-fold over


**Figure 7.** Comparison of helix C in an active CDK2 structure (red, PDB code: 1e9h) and an inactive CDK2 structure (gray, PDB code: 1hcl). In the active structure, helix C is rotated so that Glu51 can form the crucial salt bridge interaction with Lys33.

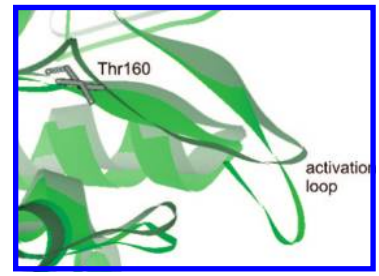


**Figure 8.** Comparison of the activation loop in a phosphorylated CDK2 structure (green, PDB code: 1e9h) and an unphosphorylated CDK2 structure (gray, PDB code: 1vyw). In the phosphorylated structure the activation loop is shifted for optimal ATP alignment and substrate stabilization. The unphosphorylated Thr160 is shown in stick form.

its analogue BMU. Nevertheless, both inhibitors induce the DFG-out conformation, and similarities between the two binding sites are detected in the clustering.[36]

- 2baq and 2bal, which also form their own cluster (cluster 5), bind highly similar inhibitors. Indeed 2baq binds Ro3201195, and 2bal binds a pyrazoloamine. The only difference between these inhibitors is the substitution of the Ro3201195 dihydroxypropane for a piperidyl group.[37]
- 1bmk, 1bl6, and 1bl7 all bind SmithKline-Beecham compounds and are found to cluster together in cluster 10.[38]
- 1wbn, 1w83, 1wbv, 1wbs, and 1wbt, which also cluster together in cluster 10, all adopt the DFG-out conformation as a result of binding a range of amide and urea analogues. In particular, 1wbs and 1wbt exhibit an extremely high binding site similarity. Inspection of their crystal structures reveals that their inhibitors have almost identical molecular geometries and exhibit very similar binding modes.[39]

*Cyclin-Dependent CDK2.* CDK2 activation is a two-step process. The first, partially activating step involves the binding of cyclin A, and the second, fully activating step involves the phosphorylation of Thr160. The initial association of cyclin A aligns ATP for phosphoryl transfer and configures the active site for catalysis.[40] The most significant conformational change caused by the binding of cyclin A is the rotation of helix C. A conserved glutamic acid residue (Glu51) at the center of helix C becomes directed toward the active site, where it forms a salt bridge interaction with a conserved lysine (Lys33) residue (Figure 7).[41] The integrity of this Lys33-Glu51 ion pair is essential for catalytic activity, since it serves to anchor the α- and $\beta$-phosphates of ATP.[5,9] Full activation results from the subsequent phosphorylation of the CDK2/cyclin A complex at Thr160. This phosphorylation shifts the activation loop to create a pocket for substrate binding (Figure 8).[40]

With a total of 80 different binding sites, CDK2 has the largest representation of all of the protein kinases in the data set. In the binding site-based classification, all CDK2 binding sites are clustered into two adjacent clusters, one of which also contains the four CDK5 binding sites (Figure 9). A more detailed analysis of the CDK2 crystal structures reveals an obvious separation based on their activation states. Indeed, all of the CDK2s that occupy cluster 1 are bound with cyclin and are in the active conformation. In addition, this cluster contains the CDK5s, which are also in the active conformation. Furthermore, this cluster is further divided into three subclusters which distinguish between the CDK5s (underlined in blue), the unphosphorylated CDK2s (underlined in yellow), and the phosphorylated CDK2s (not underlined).

Cluster 2 contains, with the exception of 1fvv, CDK2s that are all in the inactive conformation. 1fvv is in the active, unphosphorylated state, and it is unclear why it has been clustered with the inactive CDK2s. Within cluster 2, there is a clear separation of the CDK2s in terms of their source organism, with the three CDK2s from *Plasmodium falciparum* (underlined in green) branching off as an outgroup from the main group (all other CDKs in the data set are of human origin). Since the DFG-out conformation has not been observed in any inactive CDK2 structures, there is less variability in terms of the CDK2 inactive conformation, when compared with that of p38α. Indeed, the high level of similarity between inactive CDK2s is apparent from the fact that all 67 reside in the same cluster.

*Abelsone Tyrosine Kinase.* The data set contains six Abelsone tyrosine kinase (c-Abl) binding sites; three that bind pyrido[2,3.d]pyrimidine-type inhibitors (PDB codes: 1m52, 1opk, 1opl), two that bind 2-phenylaminopyrimidine-type (Gleevec-like) inhibitors (PDB codes: 1fpu, 1opj), and one that binds an aurora kinase inhibitor (PDB code: 2f4j). With the exception of 2f4j, all c-Abl binding sites were clustered into two adjacent clusters in the binding site-based classification (Figure 10). In addition to the obvious separation based on inhibitor type, the separation of these c-Abls into two different clusters reflects their differing activation states. Those with bound pyrido[2,3.d]pyrimidine-type inhibitors (P16 and P17) are in the active state, whereas those bound with Gleevec or its analogue (PRC) are in the inactive state.

Gleevec, an antileukemia drug, is known to specifically target the inactive DFG-out conformation exhibited by three different receptor tyrosine kinases: c-Abl, stem cell factor receptor c-Kit (c-Kit), and platelet-derived growth factor receptor (PDGF-R).[2] In the DFG-out conformation, the phenylalanine residue of the DFG motif is shifted by around 10 Å in relation to the DFG-in conformation that is found in active protein kinases. Gleevec, along with other so-called 'type II' inhibitors, binds in the pocket that is produced by the DFG-out conformation.[1] Part of the inhibitor molecule becomes wedged between helix C and the activation segment, therefore locking the protein in its inactive conformation.[42] Interestingly, c-Kit (PDB code: 1t46) with bound Gleevec (ligand code: STI) also resides in cluster 1, along with the c-Abl binding sites with bound Gleevec and its analogue

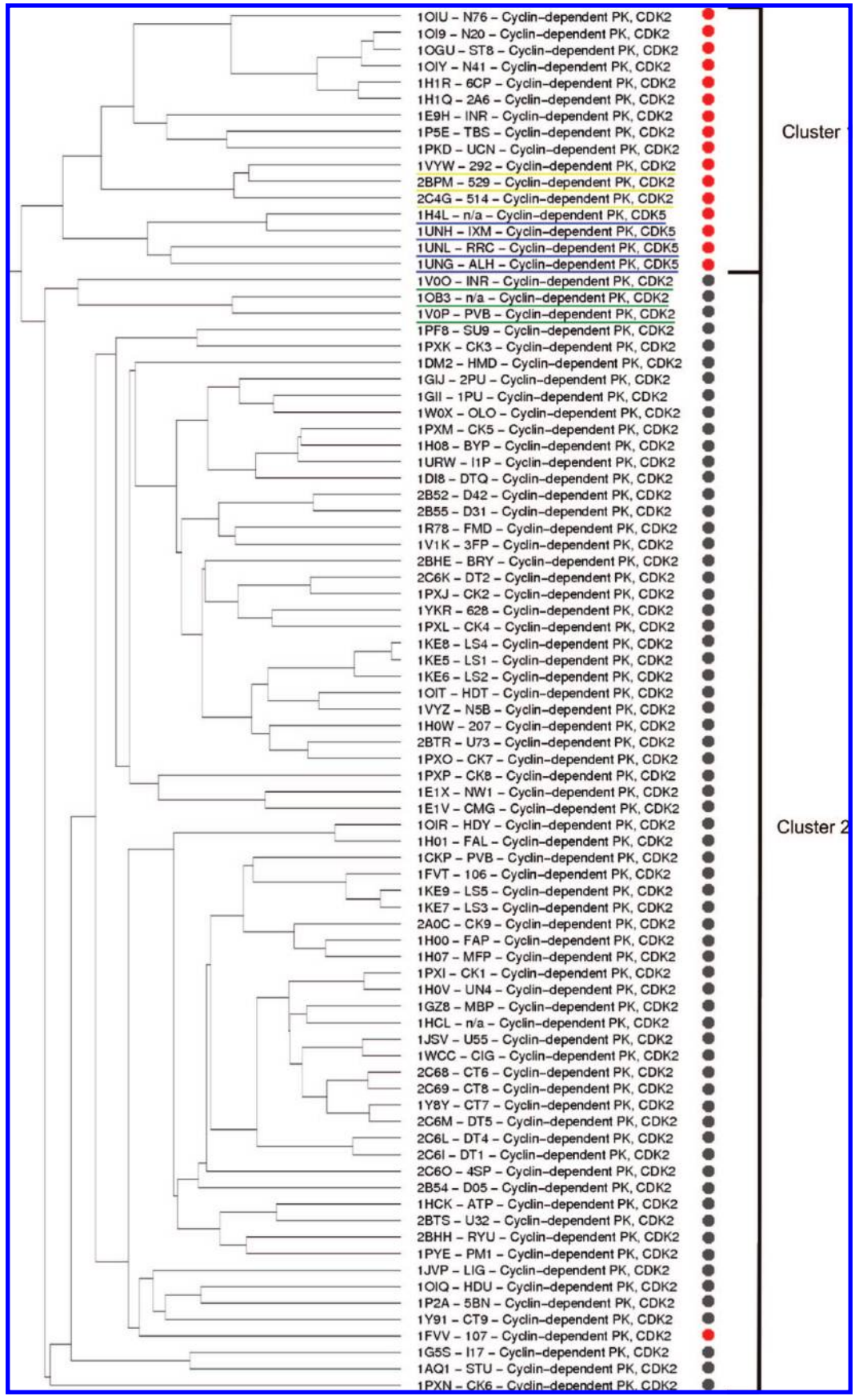


**Figure 9.** Close up view of the CDK2s taken from the binding site-based classification shown in Figure 4. Active structures have been marked with a red dot, whereas inactive structures have been marked with a gray dot. Unphosphorylated active structures have been underlined in yellow, CDK5 structures have been underlined in blue, and CDK2 structures from *Plasmodium falciparum* have been underlined in green (all other CDK structures are of human origin).

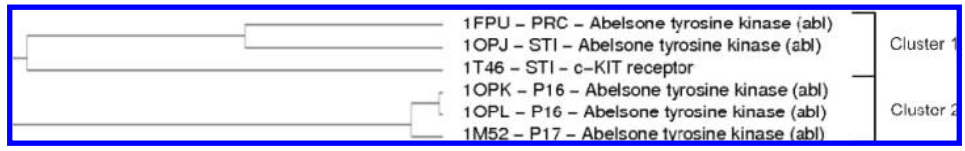


**Figure 10.** Close up view of the c-Abls taken from the binding site-based classification shown in Figure 4.

**Table 3.** Comparison of the Binding Site-Based Classification with Inhibition Data from *in Vitro* Screening Carried out by Fabian et al[a]

| drug | target | no. of sites of actives in top 5% | no. of kinase sites inhibited by drug | enrichment factor (theoretical max.) | P-value |
|---|---|---|---|---|---|
| Tarceva | EGF receptor tyrosine kinase, Erbb-1 | 9/12 | 26/239 | 6.89 (9.19) | 3.69e-08 |
| Gleevec | Abelsone tyrosine kinase (abl) | 6/12 | 20/239 | 5.98 (11.95) | 6.75e-05 |
| **SB203580** | **MAP kinase p38α** | **12**/12 | **44**/239 | **5.43** (5.43) | **4.85e-11** |
| **BIRB-796** | **MAP kinase p38α** | **11**/12 | **60**/239 | **3.65** (3.98) | **9.77e-08** |
| Gleevec | c-KIT receptor | 3/12 | 20/239 | 2.99 (11.95) | 5.54e-02 |
| **Roscovitine** | **cyclin-dependent PK, CDK5** | **12**/12 | **85**/239 | **2.81** (2.81) | **5.79e-08** |
| SP600125 | MAP kinase 8 (jnk1) | 9/12 | 99/239 | 1.81 (2.41) | 1.15e-03 |
| SP600125 | c-jun N-terminal kinase (jnk3s) | 8/12 | 99/239 | 1.61 (2.41) | 7.10e-03 |
| GW-2016 | EGF receptor tyrosine kinase, Erbb-1 | 0/12 | 2/239 | 0.00 (19.92) | – |

[a] The data in the table are ordered by descending enrichment factor. Drug-target pairs for which enrichment factors are close to or equal to their theoretical maximum are highlighted in bold.

**Table 4.** Protein Kinases with Binding Sites Ranked in the Top 5% of the Data Set in Terms of Their Similarity ($R_3$ Score) to That of the Query Kinase Binding Site, EGFR with Bound Tarceva[a]

| PDB ID | ligand ID | protein | binding site similarity ($R_3$ score) | Tarceva inhibition ($\mu$M) |
|---|---|---|---|---|
| **1m17** | **AQ4** | **EGF receptor tyrosine kinase, Erbb-1** | **1.000** | **0.0014** |
| 1y57 | MPZ | c-src protein tyrosine kinase | 0.559 | 1.9 |
| 1m14 | n/a | EGF receptor tyrosine kinase, Erbb-1 | 0.506 | 0.0014 |
| 1yol | S03 | c-src protein tyrosine kinase | 0.491 | 1.9 |
| 1qpe | PP2 | lymphocyte kinase (lck) | 0.488 | 0.53 |
| 2f4j | VX6 | Abelsone tyrosine kinase (abl) | 0.438 | 0.77 |
| 2src | ANP | c-src protein tyrosine kinase | 0.432 | 1.9 |
| 1yom | P01 | c-src protein tyrosine kinase | 0.431 | 1.9 |
| 1xbb | STI | tyrosine-protein kinase SYK | 0.425 | – |
| 3lck | n/a | lymphocyte kinase (lck) | 0.405 | 0.53 |
| 1qpc | ANP | lymphocyte kinase (lck) | 0.401 | 0.53 |
| 1phk | ATP | glycogen phosphorylase kinase (Phk), gamma-subunit | 0.395 | – |
| 1ad5 | ANP | Fl cytokine receptor | 0.391 | – |

[a] PDB code: 1m17 (highlighted in bold). The corresponding inhibition data from the Tarceva screening by Fabian et al. is also shown. A hyphen indicates that no inhibition was observed against that particular kinase.

n-[4-methyl-3-[[4-(3-pyridinyl)-2-pyrimidinyl]amino]phenyl]-3-pyridinecarboxamide (ligand code: PRC).

2f4j, the c-Abl binding site that is absent from these two clusters, represents a H396P mutant c-Abl bound with VX-680, an aurora kinase inhibitor. The mutation destabilizes the inactive conformation that is required for the binding of Gleevec, therefore rendering this c-Abl resistant to the drug. It has been crystallized in an unphosphorylated, active conformation not previously observed in c-Abl. In fact, its structure is described as closely resembling that of the active, phosphorylated lymphocyte kinase (Lck).[43] These striking similarities in conformation are reflected in the binding site-based classification, where 2f4j is clustered among the active Lcks, in addition to binding sites from several other protein kinases.

**Comparison with Experimental Inhibition Data.** We have compared the experimental interaction profiles for protein kinase inhibitors[18,29] with our binding site similarity profiles. The aim is to find out to what extent these different approaches agree. This might be valuable in relation to predicting possible cross-reactivity with other protein kinases based on binding site similarity. 240 protein kinase binding sites out of a total of 354 in our data set were represented in the experimental screening panel used by Fabian et al.[18] Of these, nine had been cocrystallized with a total of seven different inhibitors used in the *in vitro* screening. In classical *in silico* virtual screening methods such as docking, the most important metric for evaluating the performance of a method to prioritize active compounds in a database is the enrichment factor, since this measures the extent to which a data set can be enriched so that only a much smaller subset needs to be considered. In this case, we consider an inverse (protein) virtual screening protocol, identifying other binding sites that have the potential to bind a given drug based on their similarity to the site in question.

The enrichment factors were calculated by examining the top-ranked 5% of the database for each of these nine binding sites (see Methods), and the results are presented in Table 3. For example, the data used to calculate the enrichment factor for the EGF receptor tyrosine kinase (Erbb-1) binding site with bound Tarceva are shown in Table 4. This shows the top twelve most similar sites corresponding to the top 5% in the data set. In addition to finding another Erbb-1, this approach also identifies sites of c-src, Lck, and c-Abl kinases as having highly ranked site similarity while they are known actives, showing Tarceva inhibition of 0.53−1.9

**Table 5.** Comparison of the Binding Site-Based Classification with Inhibition Data from *in Vitro* Screening Carried out by Invitrogen[a]

| drug | target | no. of sites of actives in top 5% | no. of kinase sites inhibited by drug | enrichment factor (theoretical max.) | P-value |
|---|---|---|---|---|---|
| Y27632 | rho-associated protein kinase 1 | 3/15 | 8/300 | 7.50 (37.50) | 4.55e-03 |
| Gleevec | tyrosine-protein kinase SYK | 8/15 | 22/300 | 7.27 (13.64) | 5.89e-07 |
| Gleevec | Abelsone tyrosine kinase (abl) | 8/15 | 22/300 | 7.27 (13.64) | 5.89e-07 |
| Tarceva | EGF receptor tyrosine kinase, Erbb-1 | 6/15 | 19/300 | 6.32 (15.79) | 7.00e-05 |
| **H-89** | **cAMP-dependent PK, catalytic subunit** | **15**/15 | **58**/300 | **5.17** (5.17) | **2.58e-13** |
| BIRB-796 | MAP kinase p38α | 12/15 | 54/300 | 4.44 (5.56) | 1.62e-08 |
| **Roscovitine** | **cyclin-dependent PK, CDK5** | **15**/15 | **83**/300 | **3.61** (3.61) | **3.75e-11** |
| PP2 | lymphocyte kinase (lck) | 12/15 | 75/300 | 3.20 (4.00) | 5.20e-07 |
| Gleevec | c-KIT receptor | 2/15 | 22/300 | 1.82 (13.64) | 0.273 |
| BAY-439006 | B-Raf kinase | 0/15 | 8/300 | 0.00 (37.50) | - |

[a] The data in the table are ordered by descending enrichment factor. Drug-target pairs for which enrichment factors are equal to the theoretical maximum are highlighted in bold.

**Table 6.** Binding Sites with the Top Ten Consensus Similarity Scores to the Six c-Abl Binding Sites[a]

| PDB ID | ligand ID | protein | consensus score | Fabian et al. inhibition (μM) | Invitrogen inhibition (%) |
|---|---|---|---|---|---|
| **1opk** | **P16** | **Abelsone tyrosine kinase (abl)** | **0.6518** | **0.0022** | **98** |
| **1opl** | **P16** | **Abelsone tyrosine kinase (abl)** | **0.6447** | **0.0022** | **98** |
| **1m52** | **P17** | **Abelsone tyrosine kinase (abl)** | **0.6335** | **0.0022** | **98** |
| **1opj** | **STI** | **Abelsone tyrosine kinase (abl)** | **0.5062** | **0.0022** | **98** |
| **2f4j** | **VX6** | **Abelsone tyrosine kinase (abl)** | **0.5019** | **0.0022** | **98** |
| **1fpu** | **PRC** | **Abelsone tyrosine kinase (abl)** | **0.4734** | **0.0022** | **98** |
| 1y57 | MPZ | c-src protein tyrosine kinase | 0.3855 | – | 24 |
| 3lck | n/a | lymphocyte kinase (lck) | 0.3544 | 0.062 | 98 |
| 1qpe | PP2 | lymphocyte kinase (lck) | 0.3543 | 0.062 | 98 |
| 1t46 | STI | c-KIT receptor | 0.3349 | 0.83 | 45 |
| 1qpd | STU | lymphocyte kinase (lck) | 0.332 | 0.062 | 98 |
| 1ad5 | ANP | hemapoetic cell kinase Hck | 0.3268 | – | 62 |
| 1yvj | 4ST | tyrosine-protein kinase JAK3 | 0.3206 | x | 2 |
| 2b7a | IZA | tyrosine-protein kinase JAK2 | 0.3185 | – | 1 |
| 1byg | STU | carboxyl-terminal src kinase (csk) | 0.3157 | – | 43 |
| 1yol | S03 | c-src protein tyrosine kinase | 0.3147 | – | 24 |

[a] These sites are highlighted in bold. The corresponding inhibition data from Fabian et al. and Invitrogen is shown. Hyphens indicate that no inhibition was observed, and an x indicates that the protein kinase was not included in the screening.

μM. Additionally, 301 of the protein kinase binding sites in our data set were represented in the Invitrogen screening panel.[29] Ten of these binding sites were bound with eight different inhibitors used in the *in vitro* screening. The enrichment factors for each of these eight binding sites were also calculated and are shown in Table 5.

The enrichment factors presented here represent the relationship between our binding site-based classification and the results of the inhibition profiling. An enrichment factor of greater than 1.0 signifies a relative enrichment with respect to random, and the p-value represents the probability that the data are enriched by chance alone. In the Fabian et al. data set, eight out of nine enrichment factors were calculated to be greater than one, indicating a relative enrichment (Table 3). Furthermore, three out of nine binding sites achieved enrichment factors that were close or equal to their theoretical maximum (highlighted in bold). Although an enrichment factor of zero was achieved for GW-2016, this inhibitor was shown to be active against only two protein kinases in the data set. Therefore, the probability of any binding sites of actives ranking in the top 5% of the data set was significantly reduced. This is reflected in the high theoretical maximum enrichment factor (19.92) that could be achieved for this inhibitor. The Invitrogen data set (Table 5) also showed a good overall enrichment, with two out of ten binding sites achieving enrichment factors at the theoretical maximum (highlighted in bold). Again, the one binding site that was awarded an enrichment factor of zero was bound with an inhibitor that was only shown to bind eight of the binding sites in the data set.

In order to rationalize the cross-reactivity of Gleevec, a consensus scoring procedure (see Methods) was carried out in accordance with the methods used by Kuhn et al., 2007.[21] In this way, those binding sites in the data set that exhibited the highest similarity to the six c-Abl binding sites were identified. The most highly ranked binding sites were found to be those of c-src, followed by those of Lck, c-Kit, and hemapoetic cell kinase (Hck) (Table 6). Although c-src and c-Abl are closely related, Gleevec is unable to inhibit c-src since it does not adopt the required DFG-out conformation in its inactive state.[42] As mentioned previously, the observed similarity of c-Abl to c-Kit is not unexpected since c-Kit is actually cocrystallized with Gleevec.

It is interesting to note that the method has captured similarities between the binding sites of the c-Abl and Lck protein kinases. Not only did Fabian et al. show that Gleevec is able to inhibit Lck in the nanomolar range (0.062 μM), but also the Invitrogen screening revealed that it was able to inhibit Lck and c-Abl equally well (98% inhibition). As a regulator of T-cell maturation and activation, Lck is a potential target for immunosuppression. Previous studies have shown that Gleevec treatment suppresses T-cell proliferation and affects the Lck pathway.[18] Therefore, it is possible that Gleevec may be clinically useful as an immu-

nosuppressant. Interestingly, Kuhn et al. were also able to demonstrate the similarity of the c-Abl and Lck binding sites through their consensus scoring.

## DISCUSSION

We have presented a binding site-based structural classification of the protein kinases that reveals differences and similarities between protein kinases that were not evident from a sequence-based classification. The most striking initial observations were the dissimilarity of the p38α binding sites and the similarity of the chk1 binding sites to those of other protein kinases. These findings are in support of the fact that many highly selective p38α inhibitors have been developed, whereas only a few chk1 inhibitors have been reported. Indeed, these p38α inhibitors gain their selectivity from the exploitation of selectivity determinants that distinguish the p38α active site from those of other protein kinases (including other MAP kinases).

Several protein kinases were shown to exhibit a low sequence identity to one another, yet they had highly similar binding sites and therefore show potential for unexpected cross-reactivity. In several cases, cross-reactivity was confirmed by the fact that the different protein kinases were able to bind common inhibitors. Conversely, the geometric hashing method was able to identify several protein kinases that exhibited dissimilar binding sites even though they were closely related in sequence. For instance, the p38α binding sites were shown to exhibit dissimilarity to those of the other MAP kinases and even to each other.

Closer inspection of the p38α binding sites revealed that their diversity was likely to have resulted from ligand-induced conformational change. Indeed, the p38αs adopt a wide variety of conformations in response to ligand binding, with similar ligands causing similar conformational changes in the binding site. In comparison with the p38αs, a lot less variability was observed in the binding site conformations of the CDK2s. Not only was the method clearly able to distinguish active conformations from inactive conformations but also it was able to separate the active conformation according to phosphorylation state. In addition to the CDK2s, it was also able to discriminate between activation states within the c-Abls.

Comparison of our binding site-based classification with experimental inhibition data from two independent studies yielded high enrichment factors. Almost all enrichment factors showed a relative enrichment with respect to random, with several reaching the theoretical maximum. The overall strength of the enrichment factors indicated a correlation between the drug inhibition profiles determined by the *in vitro* screening and the similarity of the binding sites using our method. Additional consensus scoring of the c-Abls revealed binding site similarity to several other protein kinases, which have all, with the exception of c-src, been shown to bind the c-Abl inhibitor Gleevec with some degree of affinity.

## CONCLUSIONS

In summary, our structure-based binding site classification of the protein kinases has revealed the great diversity of different protein kinases and their differing potentials for inhibitor selectivity. Furthermore, we have demonstrated the ability to distinguish between subtle differences in binding site conformation within different protein kinases, such as activation states and ligand-induced conformational changes. Finally, we have shown that binding site similarity can be used to identify other binding sites that are likely to bind the same or similar drugs. Knowledge of such site similarities may aid the identification of off-target protein kinases for the selectivity profiling of putative lead compounds in future. To this end, binding site similarity analysis of individual protein-ligand sites can be performed using the online resource SitesBase[44] at http://www.modelling.leeds.ac.uk/sb/.

## ACKNOWLEDGMENT

This work was funded by the BBSRC in the form of a studentship for S.L.K.

**Supporting Information Available:** Information about the protein kinase data set used and the clustering validation procedure and the full binding site-based and sequence-based classifications. This material is available free of charge via the Internet at http://pubs.acs.org.

## REFERENCES AND NOTES

(1) Fedorov, O.; Sundstrom, M.; Marsden, B.; Knapp, S. Insights for the development of specific kinase inhibitors by targeted structural genomics. *Drug Discovery Today* **2007**, *12* (9−10), 365–372.
(2) Bogoyevitch, M. A.; Fairlie, D. P. A new paradigm for protein kinase inhibition: Blocking phosphorylation without directly targeting atp binding. *Drug Discovery Today* **2007**, *12* (15−16), 622–633.
(3) Johnson, L. Protein kinases and their therapeutic exploitation. *Biochem. Soc. Trans.* **2007**, *35* (Pt 1), 7–11.
(4) Scapin, G. Structural biology in drug design: Selective protein kinase inhibitors. *Drug Discovery Today* **2002**, *7* (11), 601–611.
(5) Huse, M.; Kuriyan, J. The conformational plasticity of protein kinases. *Cell* **2002**, *109* (3), 275–282.
(6) Nolen, B.; Taylor, S.; Ghosh, G. Regulation of protein kinases; controlling activity through activation segment conformation. *Mol. Cell* **2004**, *15* (5), 661–675.
(7) Kobe, B.; Kampmann, T.; Forwood, J. K.; Listwan, P.; Brinkworth, R. I. Substrate specificity of protein kinases and computational prediction of substrates. *Biochim. Biophys. Acta* **2005**, *1754* (1−2), 200–209.
(8) Engh, R. A.; Bossemeyer, D. Structural aspects of protein kinase control-role of conformational flexibility. *Pharmacol. Ther.* **2002**, *93* (2−3), 99–111.
(9) Liao, J. J. Molecular recognition of protein kinase binding pockets for design of potent and selective kinase inhibitors. *J. Med. Chem.* **2007**, *50* (3), 409–424.
(10) Vieth, M.; Sutherland, J. J.; Robertson, D. H.; Campbell, R. M. Kinomics: Characterizing the therapeutically validated kinase space. *Drug Discovery Today* **2005**, *10* (12), 839–846.
(11) Ahn, N. G.; Resing, K. A. Cell biology. Lessons in rational drug design for protein kinases. *Science* **2005**, *308* (5726), 1266–1267.
(12) Petrelli, A.; Giordano, S. From single- to multi-target drugs in cancer therapy: When aspecificity becomes an advantage. *Curr. Med. Chem.* **2008**, *15* (5), 422–432.
(13) Manning, G.; Whyte, D. B.; Martinez, R.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the human genome. *Science* **2002**, *298* (5600), 1912–1934.
(14) Cheek, S.; Ginalski, K.; Zhang, H.; Grishin, N. V. A comprehensive update of the sequence and structure classification of kinases. *BMC Struct. Biol.* **2005**, *5*, 6.
(15) Cheek, S.; Zhang, H.; Grishin, N. V. Sequence and structure classification of kinases. *J. Mol. Biol.* **2002**, *320* (4), 855–881.
(16) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697* (1−2), 243–257.
(17) Bain, J.; Plater, L.; Elliott, M.; Shpiro, N.; Hastie, C. J.; McLauchlan, H.; Klevernic, I.; Arthur, J. S.; Alessi, D. R.; Cohen, P. The selectivity of protein kinase inhibitors: A further update. *Biochem. J.* **2007**, *408* (3), 297–315.
(18) Fabian, M. A.; Biggs, W. H., 3rd; Treiber, D. K.; Atteridge, C. E.; Azimioara, M. D.; Benedetti, M. G.; Carter, T. A.; Ciceri, P.; Edeen,

P. T.; Floyd, M.; Ford, J. M.; Galvin, M.; Gerlach, J. L.; Grotzfeld, R. M.; Herrgard, S.; Insko, D. E.; Insko, M. A.; Lai, A. G.; Lelias, J. M.; Mehta, S. A.; Milanov, Z. V.; Velasco, A. M.; Wodicka, L. M.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J. A small molecule-kinase interaction map for clinical kinase inhibitors. *Nat. Biotechnol.* **2005**, *23* (3), 329–336.

(19) Fedorov, O.; Marsden, B.; Pogacic, V.; Rellos, P.; Muller, S.; Bullock, A. N.; Schwaller, J.; Sundstrom, M.; Knapp, S. A systematic interaction map of validated kinase inhibitors with ser/thr kinases. *Proc. Natl. Acad. Sci. U.S.A.* **2007**, *104* (51), 20523–20528.

(20) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28* (1), 235–242.

(21) Naumann, T.; Matter, H. Structural classification of protein kinases using 3d molecular interaction field analysis of their ligand binding sites: Target family landscapes. *J. Med. Chem.* **2002**, *45* (12), 2366–2378.

(22) Kuhn, D.; Weskamp, N.; Hullermeier, E.; Klebe, G. Functional classification of protein kinase binding sites using cavbase. *ChemMedChem* **2007**, *2* (10), 1432–1447.

(23) Brakoulias, A.; Jackson, R. M. Towards a structural classification of phosphate binding sites in protein-nucleotide complexes: An automated all-against-all structural comparison using geometric matching. *Proteins* **2004**, *56* (2), 250–260.

(24) Finn, R. D.; Tate, J.; Mistry, J.; Coggill, P. C.; Sammut, S. J.; Hotz, H. R.; Ceric, G.; Forslund, K.; Eddy, S. R.; Sonnhammer, E. L.; Bateman, A. The pfam protein families database. *Nucleic Acids Res.* **2008**, *36*, D281−288.

(25) Godden, J. W.; Xue, L.; Stahura, F. L.; Bajorath, J. Searching for molecules with similar biological activity: Analysis by fingerprint profiling. *Pac. Symp. Biocomput.* **2000**, 566–575.

(26) http://glaros.Dtc.Umn.Edu/gkhome/views/cluto (accessed on 15/11/2007).

(27) Consortium, U. The universal protein resource (uniprot). *Nucleic Acids Res.* **2008**, *36*, D190−195.

(28) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. Clustal w: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22* (22), 4673–4680.

(29) http://www.invitrogen.com/downloads/SelectScreen_Data_193.pdf (accessed on 03/03/2008).

(30) Fitzgerald, C. E.; Patel, S. B.; Becker, J. W.; Cameron, P. M.; Zaller, D.; Pikounis, V. B.; O′Keefe, S. J.; Scapin, G. Structural basis for p38alpha map kinase quinazolinone and pyridol-pyrimidine inhibitor specificity. *Nat. Struct. Biol.* **2003**, *10* (9), 764–769.

(31) Goldstein, D. M.; Alfredson, T.; Bertrand, J.; Browner, M. F.; Clifford, K.; Dalrymple, S. A.; Dunn, J.; Freire-Moar, J.; Harris, S.; Labadie, S. S.; La Fargue, J.; Lapierre, J. M.; Larrabee, S.; Li, F.; Papp, E.; McWeeney, D.; Ramesha, C.; Roberts, R.; Rotstein, D.; San Pablo, B.; Sjogren, E. B.; So, O. Y.; Talamas, F. X.; Tao, W.; Trejo, A.; Villasenor, A.; Welch, M.; Welch, T.; Weller, P.; Whiteley, P. E.; Young, K.; Zipfel, S. Discovery of s-[5-amino-1-(4-fluorophenyl)-1h-pyrazol-4-yl]-[3-(2,3-dihydroxypropoxy)ph enyl]methanone (ro3201195), an orally bioavailable and highly selective inhibitor of p38 map kinase. *J. Med. Chem.* **2006**, *49* (5), 1562–1575.

(32) Herberich, B.; Cao, G. Q.; Chakrabarti, P. P.; Falsey, J. R.; Pettus, L.; Rzasa, R. M.; Reed, A. B.; Reichelt, A.; Sham, K.; Thaman, M.; Wurz, R. P.; Xu, S.; Zhang, D.; Hsieh, F.; Lee, M. R.; Syed, R.; Li, V.; Grosfeld, D.; Plant, M. H.; Henkle, B.; Sherman, L.; Middleton, S.; Wong, L. M.; Tasker, A. S. Discovery of highly selective and potent p38 inhibitors based on a phthalazine scaffold. *J. Med. Chem.* **2008**, *51* (20), 6271–6279.

(33) Foloppe, N.; Fisher, L. M.; Francis, G.; Howes, R.; Kierstan, P.; Potter, A. Identification of a buried pocket for potent and selective inhibition of chk1: Prediction and verification. *Bioorg. Med. Chem.* **2006**, *14* (6), 1792–1804.

(34) Wood, E. R.; Truesdale, A. T.; McDonald, O. B.; Yuan, D.; Hassell, A.; Dickerson, S. H.; Ellis, B.; Pennisi, C.; Horne, E.; Lackey, K.; Alligood, K. J.; Rusnak, D. W.; Gilmer, T. M.; Shewchuk, L. A unique structure for epidermal growth factor receptor bound to gw572016 (lapatinib): Relationships among protein conformation, inhibitor off-rate, and receptor activity in tumor cells. *Cancer Res.* **2004**, *64* (18), 6652–6659.

(35) Smaill, J. B.; Baker, E. N.; Booth, R. J.; Bridges, A. J.; Dickson, J. M.; Dobrusin, E. M.; Ivanovic, I.; Kraker, A. J.; Lee, H. H.; Lunney, E. A.; Ortwine, D. F.; Palmer, B. D.; Quin, J., 3rd; Squire, C. J.; Thompson, A. M.; Denny, W. A. Synthesis and structure-activity relationships of n-6 substituted analogues of 9-hydroxy-4-phenylpyrrolo[3,4-c]carbazole-1,3(2h,6h)-diones as inhibitors of wee1 and chk1 checkpoint kinases. *Eur. J. Med. Chem.* **2007**, .

(36) Pargellis, C.; Tong, L.; Churchill, L.; Cirillo, P. F.; Gilmore, T.; Graham, A. G.; Grob, P. M.; Hickey, E. R.; Moss, N.; Pav, S.; Regan, J. Inhibition of p38 map kinase by utilizing a novel allosteric binding site. *Nat. Struct. Biol.* **2002**, *9* (4), 268–272.

(37) Sullivan, J. E.; Holdgate, G. A.; Campbell, D.; Timms, D.; Gerhardt, S.; Breed, J.; Breeze, A. L.; Bermingham, A.; Pauptit, R. A.; Norman, R. A.; Embrey, K. J.; Read, J.; VanScyoc, W. S.; Ward, W. H. Prevention of mkk6-dependent activation by binding to p38alpha map kinase. *Biochemistry* **2005**, *44* (50), 16475–16490.

(38) Wang, Z.; Canagarajah, B. J.; Boehm, J. C.; Kassisa, S.; Cobb, M. H.; Young, P. R.; Abdel-Meguid, S.; Adams, J. L.; Goldsmith, E. J. Structural basis of inhibitor selectivity in map kinases. *Structure* **1998**, *6* (9), 1117–1128.

(39) Gill, A. L.; Frederickson, M.; Cleasby, A.; Woodhead, S. J.; Carr, M. G.; Woodhead, A. J.; Walker, M. T.; Congreve, M. S.; Devine, L. A.; Tisi, D.; O′Reilly, M.; Seavers, L. C.; Davis, D. J.; Curry, J.; Anthony, R.; Padova, A.; Murray, C. W.; Carr, R. A.; Jhoti, H. Identification of novel p38alpha map kinase inhibitors using fragment-based lead generation. *J. Med. Chem.* **2005**, *48* (2), 414–426.

(40) Bartova, I.; Otyepka, M.; Kriz, Z.; Koca, J. Activation and inhibition of cyclin-dependent kinase-2 by phosphorylation; a molecular dynamics study reveals the functional importance of the glycine-rich loop. *Protein Sci.* **2004**, *13* (6), 1449–1457.

(41) Bramson, H. N.; Corona, J.; Davis, S. T.; Dickerson, S. H.; Edelstein, M.; Frye, S. V.; Gampe, R. T., Jr.; Harris, P. A.; Hassell, A.; Holmes, W. D.; Hunter, R. N.; Lackey, K. E.; Lovejoy, B.; Luzzio, M. J.; Montana, V.; Rocque, W. J.; Rusnak, D.; Shewchuk, L.; Veal, J. M.; Walker, D. H.; Kuyper, L. F. Oxindole-based inhibitors of cyclin-dependent kinase 2 (cdk2): Design, synthesis, enzymatic activities, and x-ray crystallographic analysis. *J. Med. Chem.* **2001**, *44* (25), 4339–4358.

(42) Noble, M. E.; Endicott, J. A.; Johnson, L. N. Protein kinase inhibitors: Insights into drug design from structure. *Science* **2004**, *303* (5665), 1800–1805.

(43) Young, M. A.; Shah, N. P.; Chao, L. H.; Seeliger, M.; Milanov, Z. V.; Biggs, W. H., III; Treiber, D. K.; Patel, H. K.; Zarrinkar, P. P.; Lockhart, D. J.; Sawyers, C. L.; Kuriyan, J. Structure of the kinase domain of an imatinib-resistant abl mutant in complex with the aurora kinase inhibitor vx-680. *Cancer Res.* **2006**, *66* (2), 1007–1014.

(44) Gold, N. D.; Jackson, R. M. SitesBase: A database for structure-based protein-ligand binding site comparisons. *Nucleic Acids Res.* **2006**, *34* (database issue), D231–234.

CI800289Y