

CORES: An Automated Method for Generating Three-Dimensional Models of Protein/Ligand Complexes

Brian J. Hare,* W. Patrick Walters, Paul R. Caron, and Guy W. Bemis

Vertex Pharmaceuticals, 130 Waverly Street, Cambridge, Massachusetts 02139

Received February 3, 2004

We describe a new, automated method for building 3D models of small-molecule ligands complexed with proteins. Modeling templates are constructed from frameworks (i.e., ring systems and linkers) of ligands extracted from 3D structures of ligands complexed with proteins that are structurally related to the target protein. These templates are typically substructures of the target ligand and are used to build models that constrain the ligand's conformation and binding orientation in the active site of the target protein. The practical utility of the method is shown by demonstrating that most ligands containing related frameworks bind protein kinases in the same orientation. Moreover, models for 15 of 19 cdk2/ligand complexes in the protein data bank built using our method deviate from the X-ray structure by less than 2 Å (rms). Finally, we show that over 70% of small-molecule protein kinase inhibitors published in *J. Med. Chem.* since 1993 can be modeled using a template extracted from a 3D protein kinase structure in the protein data bank.

Introduction

Structure-based drug design often relies on high-quality three-dimensional (3D) models of small-molecule inhibitors bound to drug targets. Three-dimensional models are used throughout the drug design process for visual analysis and molecular docking. Since experimental determination of 3D structures is still slow relative to the identification of new inhibitor molecules, development of accurate computational methods for model building is extremely important.

Automated modeling of small-molecule structures in protein complexes is typically performed using molecular docking.^{1–4} Docking attempts to model the 3D structure of a protein/ligand complex by optimizing the conformation and orientation of the ligand in the protein binding site using all of the conformational, translational, and rotational degrees of freedom available to the ligand. Typically, docking does not directly utilize information about the conformation and orientation of key binding elements in the target ligand from 3D structures of complexes closely related to the target complex (i.e., a ligand similar to the target ligand bound to a protein with active site similar to the target protein active site). Although a recently described method uses the binding orientation of a molecular scaffold from one X-ray structure to generate *in silico* inhibitor selectivity profiles for a set of related enzymes,⁵ automated selection of a crystallographic scaffold pose for use in modeling has not been reported.

In contrast to molecular docking, information from related 3D structures is used routinely and in an automated fashion for modeling 3D protein structures. For example, comparative homology modeling approximates a 3D protein structure based on the structures of one or more related proteins.^{6,7} Comparative modeling software packages^{8,9} automatically identify related 3D

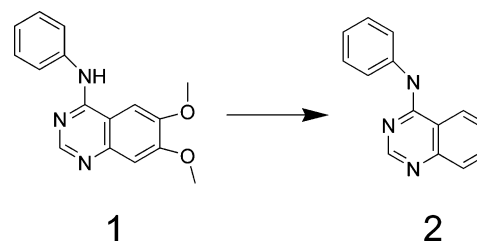


Figure 1. Generation of molecular frameworks. Molecule **1** from pdb code 1di8¹³ is reduced to the molecular framework **2** by removing the 6,7-dimethoxy moieties.

structures and extract protein backbones from these structures for use as protein modeling templates.

Just as protein backbones can serve as templates for modeling protein structures, molecular frameworks can serve as templates for modeling ligands in ligand/protein complexes. Frameworks are the union of ring systems and linkers in a molecule (see Figure 1). Each molecule has a single framework. For example, the framework for phenyl *p*-tolylamine is diphenylamine. Frameworks are useful as templates for 3D model building because large molecular databases often have a relatively small number of common frameworks. For example, 42 molecular frameworks accounted for a quarter of 5120 drugs analyzed in the Comprehensive Medicinal Chemistry (CMC) database.¹⁰ Furthermore, ligand frameworks frequently contain key protein-recognition elements (e.g., hydrogen-bonding atoms and hydrophobic moieties) that determine ligand binding orientation in protein/ligand complexes. For example, two out of three hydrogen bonds typically formed between the adenosine moiety in ATP and the hinge region of protein kinases are formed with atoms in the ATP framework. Frameworks are also easy to manipulate computationally. Thus, reduction of molecular databases to frameworks and selection of appropriate frameworks for model building is easily automated.

* To whom correspondence should be addressed. Phone: 617-444-6595. Fax: 617-444-6680. E-mail: brian_hare@vrtx.com.

Here, we describe a new method for building 3D models of protein/ligand complexes. We use molecular frameworks, selected from a database of experimental 3D structures of ligand/protein complexes, as modeling scaffolds. To evaluate the utility of frameworks as ligand templates, we analyze complexes in the protein data bank containing a small-molecule ligand bound in the ATP site of a protein kinase. We analyze protein kinases because many are thought to be good drug targets¹¹ and a significant number of chemically diverse kinase inhibitors and X-ray structures of inhibitor/protein kinase complexes are publicly disclosed. Comparing identical frameworks and frameworks sharing common substructures, we show that no framework occurs in more than four distinct binding orientations in our data set. Moreover, ligands that share a common framework containing three or more rings usually all bind in the same orientation. We build models of cdk2/ligand complexes with X-ray structures in the Protein Data Bank¹² (pdb) and show that 15 of the 19 structures can be modeled accurately by our method. Finally, we demonstrate the practical utility of the method using a purchased database of 377 protein kinase inhibitors published in the *Journal of Medicinal Chemistry* between 1993 and 2002. While only 10 molecules in this database are identical to compounds with an X-ray structure in complex with a protein kinase in the pdb, we show that 72% share a framework and therefore can be modeled using our method.

Methods

All software was written at Vertex Pharmaceuticals, Inc. in Python, Perl, or C++ unless otherwise noted. Routines that require molecular representation use the Python or C++ interface to the OECHEM library (OpenEye Scientific Software, Santa Fe, NM 87507).

X-ray Structures. We used FASTA¹⁴ to identify X-ray structures in the pdb with sequences homologous to the kinase domain of pka α using a cutoff value of 3. Because we used a high cutoff value, the choice of a reference kinase sequence does not affect the results. Only structures containing a ligand that binds to the ATP pocket of the kinase were included in the analysis. For pdb files containing multiple structures of the same kinase domain with different chain names, only the first chain containing the kinase domain was included in the analysis. As shown in Figure 2, the X-ray structures were aligned in a common coordinate frame by superimposing backbone atoms (N, CA, and C) of residues corresponding to 142–149 in the jnk3 hinge region¹⁵ onto the jnk3 reference structure (pdb code 1jnk¹⁶) using the McLachlan algorithm¹⁷ as implemented in the program ProFit (A. C. R. Martin, <http://www.bioinf.org.uk/software/profit/>).

Separate files for ligand and protein atoms were extracted from each aligned pdb file. A SMILES string was obtained for each ligand by converting the IUPAC name in the HETNAM record of the pdb file to SMILES using Chemdraw (CambridgeSoft, Cambridge, MA 02140) with manual error checking. The SMILES string and pdb coordinates were then used to create an MDL mol file (MDL Information Systems, San Leandro, CA 94577). A framework library was created by reducing the molecules to frameworks using the method de-

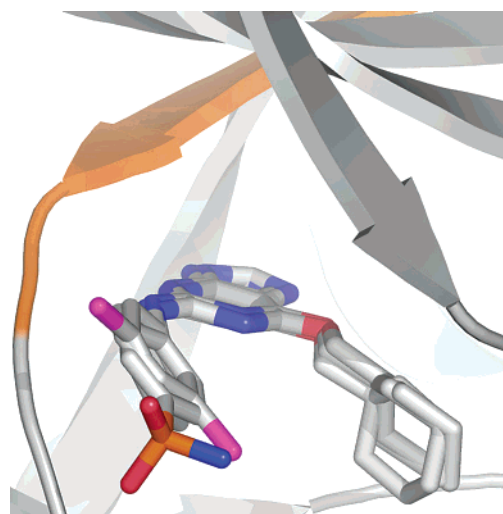


Figure 2. Align 3D structures. Shown are three X-ray structures of protein kinase inhibitors with identical frameworks (pdb codes 1h1q,¹⁸ 1h1r,¹⁸ 1h1s¹⁸). The X-ray structures of the three inhibitors are in complex with the protein kinase cdk2, and the cdk2 structure from 1h1q¹⁸ is shown (ribbon drawing). The complexes were aligned in a common reference frame by superimposing the backbone atoms (shown in orange) for each cdk2 protein structure to the corresponding atoms in a jnk3 X-ray structure (pdb code 1jnk¹⁶). Figure was prepared using PyMOL.¹⁹

scribed by Bemis and Murcko¹⁰ (Figure 1) except that here we include in the molecular framework carbonyl oxygen atoms directly connected to framework atoms. We include these atoms because they are rigid and because they sometimes make important interactions with proteins.

Binding Mode Analysis. From the library containing ligand frameworks from protein kinase X-ray structures, we identified sets of identical frameworks (illustrated in Figure 3a) and sets of related frameworks (illustrated in Figure 3b). To construct sets of related frameworks, each framework in the library was used in turn as a query ligand. For each query ligand, a framework set was created that contained the query ligand and all of the other ligands in the framework library that have the query ligand as a substructure (illustrated in Figure 3b). Only framework sets containing the query ligand and at least one other ligand from the framework library were saved for further analysis.

We determined the number of binding orientations in protein kinase ATP sites for each framework set by first calculating the root-mean-square (rms) distance between corresponding framework atoms in each pair of molecules within the set. For sets of related frameworks, rms distances were calculated using only the atoms in the common framework substructure. We then clustered the molecules in each set using the single-linkage method²² with a cutoff of 1.5 Å. Each separate cluster identified by this procedure was counted as a distinct binding mode.

Model Building. All computations were carried out on an Intel Xeon processor (2.20 GHz) with a cache size of 512 kB. We constructed a template list using the library containing ligand frameworks from protein kinase X-ray structures. The model building procedure CORES (complexes restricted by experimental structures) is described step-by-step below.

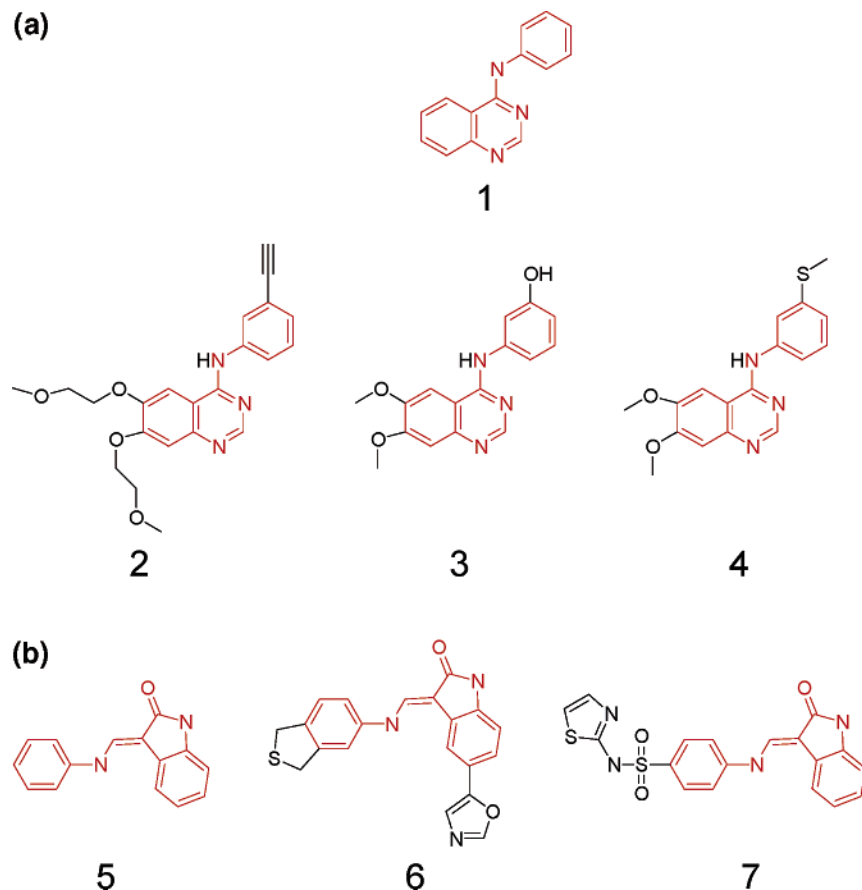


Figure 3. (a) Example of ligands from protein kinase X-ray structures with identical frameworks. **2** (pdb code 1m17²⁰), **3** (pdb code 1di8¹³), and **4** (pdb code 1di9¹³) all share the framework **1**. (b) A set of related frameworks. All frameworks are from ligands in protein kinase X-ray structures: **5** (pdb code 1ke5²¹), **6** (pdb code 1ke7²¹), **7** (pdb code 1ke8²¹).

1. The first step in the restricted docking process is the identification of template frameworks, **T**, that are substructures of the molecule, **M**, to be modeled. We identify templates in two ways. First, we perform a subgraph match of each molecule, **T**, in the framework library with **M** (Figure 4a). Frameworks with successful subgraph match are added to the list of suitable templates. Second, we perform a subgraph match of the framework of **M** with each molecule, **T**, in the framework library (Figure 4b). For each match, a template containing the atoms in the subgraph match is created and added to the list of suitable templates.

2. Each suitable template identified in step 1 is used to define a set of fixed and a set of flexible bonds. Any rotatable bond in **M** that maps to a bond in **T** is marked as fixed, and the dihedral in **M** is set to the value observed in **T**. This process is illustrated in Figure 4c. The template, shown in red, contains two rotatable bonds with dihedrals of 175° and 139°. The corresponding dihedrals in the molecule **M** to be docked are set to the values observed in **T**. These bonds are then marked as fixed and are not searched in the third step. All remaining dihedrals are marked as flexible and searched in step 3.

3. A conformational search of the dihedrals marked as flexible in the previous step is then performed to generate an ensemble of low-energy conformers. The conformational search is carried out using the program Omega (OpenEye Scientific Software, Santa Fe, NM 87507). Omega performs a systematic search over a set

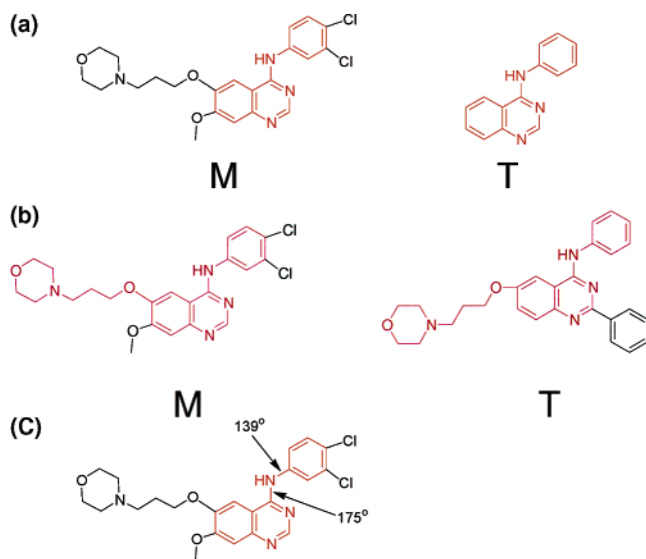


Figure 4. (a) The framework **T** is a substructure of the molecule **M** to be modeled. Atoms in the subgraph that are in both **T** and **M** are shown in red. (b) The framework of **M** is a substructure of **T**. Atoms in the subgraph that are in both the framework of **M** and **T** are shown in red. (c) Rotatable bonds in **M** that map to a bond in the template **T** from (a). Values shown for the dihedrals are from **2** (pdb code 1m17²⁰).

of discrete values for each dihedral marked as flexible in the second step. The default torsion file from OpenEye was used. The energy of each conformer is determined using a simplified force field.

Three criteria are used to limit the set of conformers generated by Omega. A maximum of 50 conformers are retained, all having a strain energy within 10 kcal of the lowest energy conformation retained. To avoid docking redundant conformers, any conformer having an rms fit of less than 0.6 Å to another conformer is removed.

4. In the final step, each conformer of the molecule **M**, to be docked, is superimposed on the template **T**. Following the superposition, the position of **M** is optimized using rigid body minimization against ChemScore.^{23,24} At the completion of the minimization, the rms displacement of the atoms in **M** corresponding to **T** from the original position of **T** is measured.

Protein Kinase Inhibitors. Inhibitors in a database of compounds published in *J. Med. Chem.* between 1993 and 2002 (GVK, Boston, MA 02109, <http://www.gvkbio.com>) that are active ($IC_{50} < 1 \mu\text{m}$) against pka, erk, cdk, p38, pdgfr, kit, or src were selected. We chose compounds active against this subset of kinases because visual inspection of the database indicated that these were the most common assayed kinases for compounds contained in the database. Compounds with peptide backbones (identified visually) and compounds with frameworks containing fewer than 7 atoms were removed, leaving a total of 377 unique inhibitors.

Results

Framework Binding Modes. We analyzed the library containing ligand frameworks from protein kinase X-ray structures. A total of 51 unique ligand frameworks are extracted from the 117 protein kinase/ligand complexes in the Protein Data Bank. One of these frameworks, 9-(tetrahydrofuran-2-yl)-9H-purine, is the framework for ATP. It is represented 51 times and always binds to protein kinases in the same orientation, so it was excluded from further analysis. The 50 remaining unique ligand frameworks are shown in Table 1, together with the pdb codes of the X-ray structures containing each ligand. Among the 50 frameworks, 14 are represented more than once in the data set (see Figure 3a for an example). A total of 33 complexes contain these 14 frameworks and 7 of the frameworks are found in complexes with more than one protein kinase.

Figure 5a is a histogram showing the distribution of number of binding modes for the 14 sets of identical frameworks. The results for different size frameworks are shown separately. The analysis reveals that the majority of the frameworks (78%) are found in a single orientation.

We extend the analysis to sets of related frameworks. We obtained 9 sets of related frameworks containing frameworks from a total of 39 unique ligands (see Figure 3b for an example). Among the 9 sets are 6 containing complexes between two or more distinct protein kinases.

A histogram showing the distribution of the number of binding modes for the 9 sets of frameworks is shown in Figure 5b. Results for different size frameworks are shown separately. A majority (55%) of the ligand sets bind in a single orientation. The common frameworks for many of the sets are small. Most contain only two rings. In contrast to the larger frameworks, which

usually bind in only one orientation, the number of binding modes for the smaller frameworks are evenly distributed between 1 and 4. The number of binding modes for each framework is shown in Table 1.

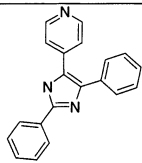
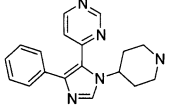
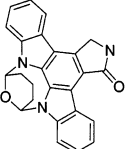
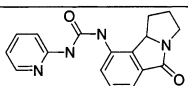
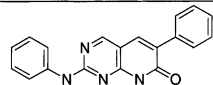

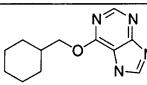
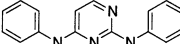
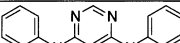
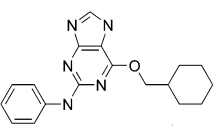
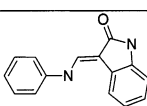
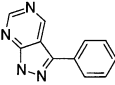
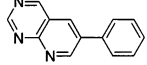
3D Models. We built and evaluated models of 21 cdk2/ligand complexes with X-ray structures in the pdb. The 21 complexes were chosen from the 32 cdk2/ligand complexes in the pdb because of the availability of suitable ligand templates for these complexes in our framework library. Among the other 11 cdk2 complexes, 9 complexes contain ligands with unique frameworks and 2 structures contain staurosporin bound to cdk2. Since model building using an identical ligand as a template is trivial, the staurosporin structures were omitted.

For each of the 21 complexes, we built multiple models as described in Methods and used the procedure described below to select a final model to compare with the X-ray structure. Among the multiple models, we first eliminated models built using any template other than the one with the most rings. We favor templates containing more rings because they usually bind in fewer discrete orientations (Figure 5) and find that models built using larger templates are more accurate (data not shown). Among the remaining models, we then selected as the final model the one with the smallest ligand displacement during rigid body minimization with the empirical scoring function ChemScore. We also evaluated ChemScore as a criterion for model selection both before and after rigid body minimization but found ligand displacement performed better (data not shown). We built all of the models using a single cdk2 X-ray structure (pdb code 1gz8²⁵), chosen because it has the highest resolution (1.3 Å) among human cdk2 X-ray structures in the Protein Data Bank.

Comparisons of the final models with corresponding X-ray structures are shown in Tables 2 and 3. We modeled 15 of the ligand complexes accurately (rms deviation less than 2.0 Å from the X-ray structure). Accurate models are distinguished by small (<1.5 Å) ligand displacement during the rigid body minimization step of model building. Ligand displacement for all of the accurate models is less than 1.5 Å and is 1.0 Å or less for 13 out of the 15 accurate models.

Six models deviated more than 2.0 Å from the X-ray structure of the complex. Two of these (pdb codes 1gij²⁶ and 1p5e²⁷) are easily filtered by large (>1.5 Å) ligand displacement during rigid body minimization. A third (pdb code 1ckp²⁸) was modeled using the framework from the ligand in pdb code 1gz8.²⁵ The ligands in pdb codes 1ckp²⁸ and 1gz8²⁵ bind in different orientations (see Figure 6), and modeling 1ckp²⁸ using the framework from 1gz8²⁵ results in an inaccurate model. No other framework in the database is a suitable modeling template for 1ckp.²⁸ The other three models deviating more than 2.0 Å from the X-ray structure were modeled using templates with the correct binding orientation (i.e., in the same framework cluster). In all three of these cases, the ligands extend out of the kinase active site and into the solvent. The positions of the ligand atoms contacting protein active site atoms are very similar in the X-ray structures and models (rms deviations of 1.1, 0.6, and 1.2 Å for pdb codes 1h06,²⁹ 1ke8,²¹ and 1g5s,³⁰

Table 1. Frameworks from Protein Kinase Inhibitors in the Protein Data Bank

Framework	PDB Codes	kinases	Number of binding modes ^a
	1a9u, 1pme	p38 α , erk2	1
	1b17, 3erk	p38 α , erk2	1
	1a41, 1byg, 1nvq, 1nvr, 1pkd, 1qpd, 1qpi, 1stc	cdk2, csk, chk-1, lck, pka α	1
	1gih, 1gii	cdk2,	1
	1m52, 1opk, 1opl	abl1,	1
	1yds, 2csn,	pka α , cki- γ 3	1
	1h1p	cdk2	1
	1h01, 1h08	cdk2	1
	1h00, 1h06, 1h07	cdk2	1
	1h1q, 1h1r, 1h1s	cdk2	1
	1ke5, 1ke9	cdk2	1
	11qcf, 1qpe	hck, lck	1
	2fgi,	fgfr1	1

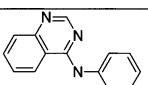

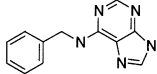
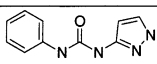
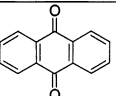

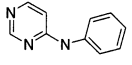
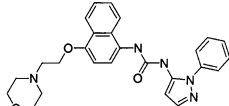
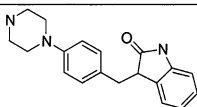
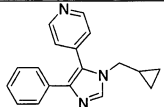
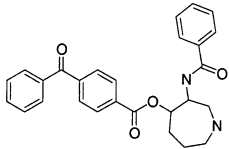
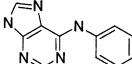
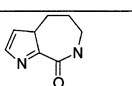
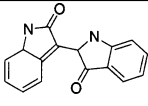
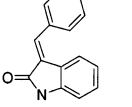
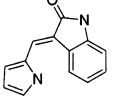
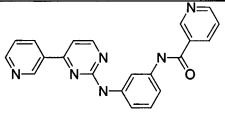
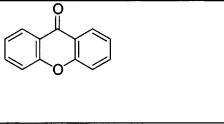
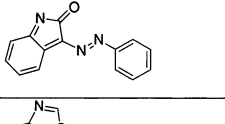
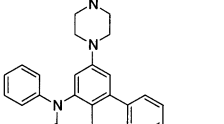
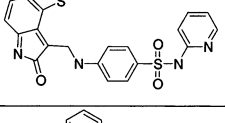
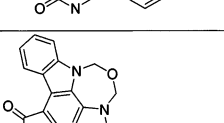
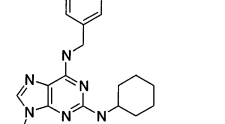
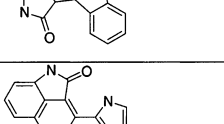
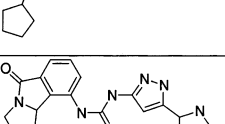
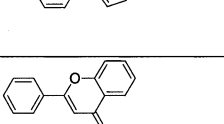
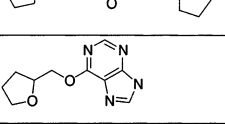
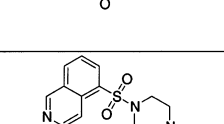
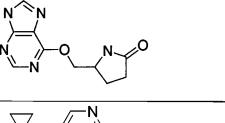
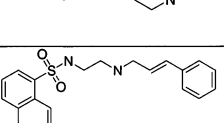
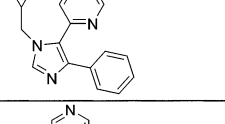
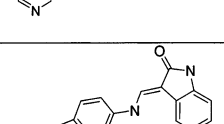
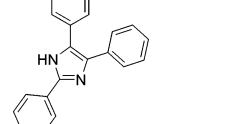
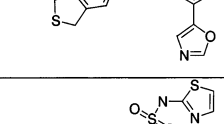
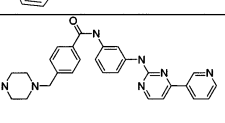
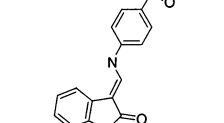
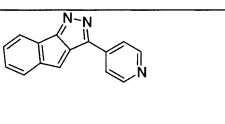
	1di8, 1di9, 1m17	cdk2, p38 α , egfr	2
	1j91, 1p5e	ckii α	2
	4erk	erk2	2
	1kv1	p38 α	2
	1f0q, 1m2p, 1m2r	ckii α	3
	1jpa, 1gz8	ephb2, cdk2	3
	1jsv	cdk2	4
	1kv2	p38 α	ND
	1agw	fgfr1	ND
	1bl6	p38 α	ND
	1bx6	pka α	ND
	1ckp	cdk2	ND
	1dm2	cdk2	ND
	1e9h	cdk2	ND
	1eh4	casein kinase-1 (yeast)	ND
	1fgi	fgfr1	ND

Table 1 (Continued)

	lfpv	abl1	ND		1m2q	ckii α	ND
	lfvt	cdk2	ND		1m7q	p38 α	ND
	lfvv	cdk2	ND		1nvs	chk-1	ND
	lg5s	cdk2	ND		1p2a	cdk2	ND
	lgij	cdk2	ND		2hck	hck	ND
	lh0u	cdk2	ND		1ydr	pkac α	ND
	lh0v	cdk2	ND		1ydt	pkac α	ND
	lbmk	p38 α	ND		1ke7	cdk2	ND
	lian	p38 α	ND		1ke8	cdk2	ND
	liep, 1opj	abl1	ND		1ke6	cdk2	ND
	ljvp	cdk2	ND				

^a ND indicates that the framework has only appeared once so far in protein kinase X-ray structures in the pdb.

respectively). The relatively high rms deviations for these models result from different orientations for moieties that protrude into the solvent and away from the active site. Thus, these models are still quite useful for analysis of ligand binding within the active site, despite having relatively high overall rms deviation from the X-ray structure. By exclusion of the two models eliminated by the ligand displacement filter, 15 of 19 models (79%) that were built using our method deviated by less than 2.0 Å (rms) from the X-ray structures.

Protein Kinase Inhibitors. In addition to being accurate, model building techniques must be broadly applicable in order to be useful. Therefore, we searched

for templates in our framework library that could be used to model molecules in a database of 377 protein kinase inhibitors published in the *Journal of Medicinal Chemistry* (1993–2002). Only 10 molecules in the *J. Med. Chem.* database are identical to ligands in protein kinase X-ray structures (Table 4). However, the frameworks of 85 inhibitors, or 23%, are identical to the framework of a ligand in a protein kinase X-ray structure. A total of 9 distinct frameworks were matched. The framework matched most often is shown in Figure 8 (8). The X-ray structure of the compound containing 8 is a complex with Abl tyrosine kinase (pdb code 1m52³¹). The frameworks for a total of 27 different

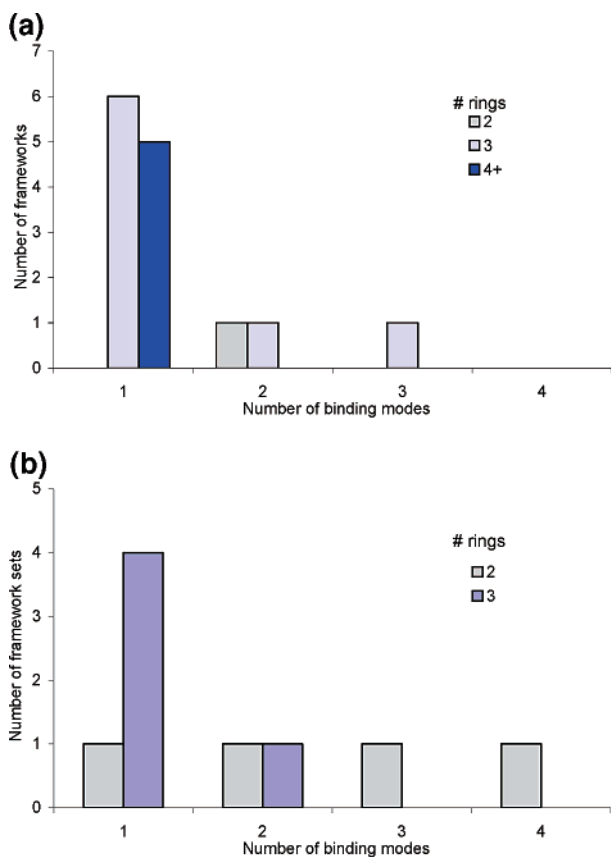


Figure 5. Histogram showing number of binding modes for frameworks from protein kinase/inhibitor complexes in the protein data bank. Frameworks with two, three, and four or more rings are plotted separately. (a) Identical frameworks are compared. (b) Related frameworks are compared.

Table 2. 3D Models of cdk2/Ligand Complexes

ligand (pdb code)	ligand template (pdb code)	ligand model ^a vs ligand X-ray structure (rms, Å)	rms displacement during minimization, Å	no. of template rings
1ckp	1gz8	3.9	0.13	2
1di8	1di9	1.8	1.3	3
1e1v	1h1p	1.4	1.3	3
1g5s	4erk	2.1	1.0	3
1gij	1kv1	12.5	3.8	2
1h01	1h08	1.4	0.7	3
1h06	1h07	2.8	0.4	3
1h07	1h06	1.7	0.7	3
1h08	1h01	1.9	0.5	3
1h0u	1gz8	1.5	0.2	2
1h1p	1gz8	2.0	0.3	2
1h1q	1h1r	0.8	0.4	4
1h1r	1h1q	0.8	0.4	4
1h1s	1h1r	1.0	0.6	4
1ke5	1ke9	1.3	0.4	3
1ke7	1ke5	1.2	1.0	3
1ke8	1ke9	2.8	0.4	3
1ke9	1ke5	1.3	0.8	3
1p5e	1j91	4.0	2.4	2
1e1x	1h1s	1.5	0.31	2
1jsv	1h07	1.7	0.8	2

^a Following rigid body minimization.

inhibitors in the *J. Med. Chem.* database are identical to **8**. These inhibitors are broadly active against tyrosine kinases.³²

An additional 117 protein kinase inhibitors from the *J. Med. Chem.* database have the ligand framework

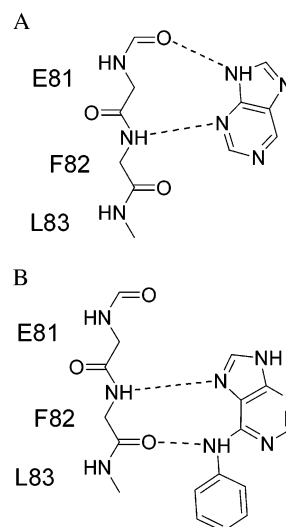


Figure 6. Distinct framework hydrogen bond interactions with the hinge for the ligands in pdb code 1gz8²⁵ (A) and 1ckp²⁸. (B). Ligand hydrogen atoms involved in hydrogen bonds are shown.

Table 3. Distribution of cdk2/Ligand Models with Respect to Ligand Displacement and rms Deviation of the Model from the X-ray Structure

ligand displacement, Å	rms vs X-ray structure, Å	
	≤ 2	> 2
≤ 1.5	15	4
> 1.5	0	2

Table 4. Distribution of Template Types from the Protein Data Bank for Modeling Protein Kinase Inhibitors in the *J. Med. Chem.* Database

template type	no. of compds	no. of distinct pdb templates
identical molecule	10	10
identical framework	85	9
substructure ^a	117	11
substructure ^b	59	6

^a A molecule in the template library is a substructure of the inhibitor to be modeled. ^b Framework of the inhibitor to be modeled is a substructure of a molecule in the template library.

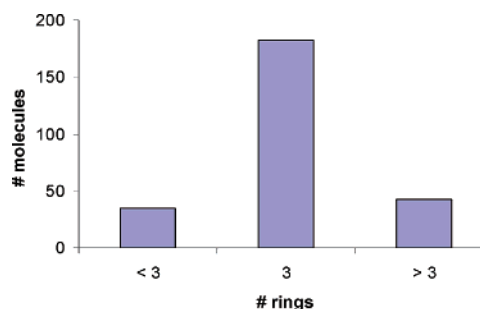


Figure 7. Distribution of the number of rings in the common frameworks (from Figure 6a).

from a protein kinase X-ray structure as a substructure (e.g., Figure 3b). Keeping only the largest among the framework substructures for each of these inhibitors, we find a total of 11 distinct ligand frameworks. The most common among these 11 frameworks (**9**) is a substructure of 50 inhibitors. **9** is the framework for an inhibitor of the fibroblast growth factor receptor tyrosine kinase domain (pdb code 2fgi³³) and is a substructure of the framework matched most often in the identical

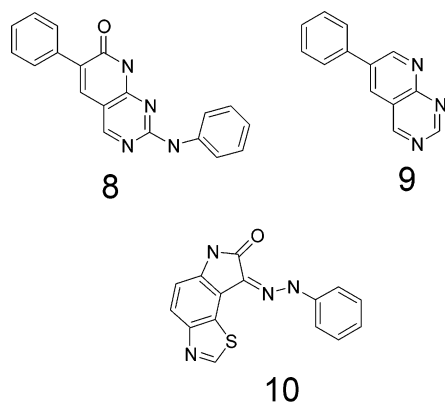


Figure 8. Ligand frameworks extracted from protein kinase complexes in the pdb that are the most common modeling templates for kinase inhibitors from *J. Med. Chem.* database.

framework search (**8**). Twenty-three distinct ligand frameworks in the *J. Med. Chem.* database were matched by **9** in the substructure search.

The frameworks for an additional 59 ligands, or 16% of the inhibitor database, are themselves a substructure of 6 different ligand frameworks in protein kinase X-ray structures (see Figure 4b). Frameworks from 29 of these inhibitors are substructures of **10**. These 29 molecules are inhibitors of cyclin-dependent kinases.³⁴

In total, 72% of the protein kinase inhibitors analyzed can be modeled using our method. The remaining inhibitors cannot currently be modeled because a suitable modeling template is not yet present in the database. However, as kinase cocomplex X-ray structures continue to be determined, the number of inhibitors that can be modeled using our method will continue to increase. Moreover, a focused effort to determine representative cocomplex structures of molecules from each chemical class of known inhibitors could efficiently increase X-ray structure coverage of known inhibitor space. In the interim, molecular docking will continue to be useful in order to identify likely binding modes of inhibitors that cannot be predicted using our method.

The protein kinase inhibitor cocomplex models should be quite accurate. Figure 7 shows that templates with three or more rings are found for a large majority (87%) of the inhibitors that can be modeled using our method. As we showed (Figure 5), larger framework templates are less likely to bind in multiple orientations than smaller frameworks and therefore should produce more accurate models. The improved accuracy of the models built using larger frameworks is demonstrated in Table 2, where the average rms deviation from X-ray structures for models built using two rings, three rings, and four rings is 3.87, 1.79, and 0.87 Å, respectively. The fraction of the public inhibitor database that can be modeled using framework templates containing three or more rings is greater than the corresponding fraction in the cdk2 test set (Table 2). Therefore, we expect that the accuracy of modeling the public inhibitors will exceed the 79% accuracy rate obtained for the cdk2 inhibitors.

Discussion

We propose a new, automated method for building 3D models of small-molecule ligands bound to protein targets. Our method uses ligand frameworks from X-ray

structures of protein/ligand complexes structurally related to the target complex as ligand templates for model building. The method extends and automates the process modelers and chemists already use to hypothesize the binding mode for an inhibitor based on X-ray structures of related complexes. To evaluate the potential to use this method for high-throughput model building, we analyzed the public-domain kinase X-ray structures and a data set of known kinase inhibitors.

For our method to be useful, ligands containing the same framework must bind related proteins in discrete orientations. Moreover, we would like to be able to predict in advance whether a ligand can be modeled accurately with existing X-ray structural data. This requires knowing the number of discrete orientations with which ligands containing the same framework can be expected to bind.

Parts a and b of Figure 4 show that the size of a framework indicates whether it is likely to bind protein kinases in multiple orientations. By combination of the results in parts a and b of Figure 4, 82% of the clusters with a core framework containing three or more rings bind in a single orientation. No framework containing four or more rings binds in more than one orientation. In contrast, core frameworks containing only two rings bind using a single orientation in only 20% of the framework sets.

Interestingly, ATP contains three rings, and molecules containing the ATP framework (e.g., ATP analogues and adenosine) all bind in the same orientation in complex with protein kinases. More generally, endogenous cofactors and substrates may have to bind in a single orientation in order to avoid nonproductive orientations of these ligands that might inhibit biological pathways. We may therefore be able to use natural ligands to predict the size of molecular templates that will likely adopt unique binding orientations in a protein binding pocket.

It is clearly preferable to use larger frameworks as modeling templates. However, sometimes only smaller templates may be available. It is therefore useful to identify models built using template ligands in the proper orientation. We find that because models built using templates in the proper orientation are usually near an energy minimum, ligand displacement during rigid body minimization is often large for inaccurate models. Additional filter functions such as ligand strain energy may also eliminate inaccurate models.

We found that, by use of ligand displacement as a filter, only 4 of 19 models built using our method deviated from the X-ray structure by more than 2.0 Å. In three of these cases, the difference was due primarily to ligand atoms outside the protein active site, suggesting that the quality of our final models could be improved by more rigorous minimization of solvent-exposed ligand fragments. The remaining inaccurate model was built using a framework binding in an orientation that differs substantially from the binding orientation of the target ligand. In this case, a framework with the correct binding mode was not present in the database. As more X-ray structures of kinase/inhibitor complexes are determined, errors of this type will become even less common. Overall, our accuracy rate of 79% is already in line with validation results

for the GOLD molecular docking software showing accurate models for 6 out of 7 kinases (86%) in their data set.⁴

Since we start with fewer molecular poses, our method is faster than molecular docking. We built all the models for 21 complexes in about 90 s, compared to typical run times of 1–5 min per compound for molecular docking with conformational flexibility. Our method can be made even faster by using only the largest suitable templates to build models.

Our method also requires less sophisticated algorithms for pose generation, minimization, and scoring. Moreover, since discrimination among models built using different framework orientations is based on the rms distance of the initial molecular pose from the nearest local minimum rather than on a score related to the energy of the complex, our method is less likely to be sensitive to small protein conformational changes. Indeed, all of the models shown in Table 3 were built using a single protein X-ray structure.

The use of frameworks for modeling these complexes has a number of limitations that will be addressed in future work. First, information from acyclic groups is lost even when it is a critical protein recognition feature. For example, the ligands in pdb codes 1m2q³⁵ and 1m2r³⁵ have similar frameworks but bind to the protein kinase cKII in different orientations because of different interactions between acyclic ligand atoms and the protein kinase hinge. Second, peptide-based ligands are difficult to model because peptide and peptidomimetic backbones can be mapped onto one another in multiple orientations. Finally, simple heteroatom substitutions prevent template matches (e.g., pyridine will not be mapped onto pyrimidine even if protein recognition requires only the pyridine nitrogen). With heteroatom substitution, an additional six compounds in the public inhibitor data set have frameworks that match exactly to frameworks from public kinase X-ray structures.

We believe that our method will gain increasing favor as the number and diversity of 3D structures of proteins complexed with small molecules increases. More than 70% of protein kinase inhibitors in a database of public domain protein kinase inhibitors can already be modeled using our method (Figure 6). The method is readily extendable to modeling small molecules bound to the binding sites for ATP, cofactor, or substrates in other protein families (e.g., lipid kinases, inosine monophosphate dehydrogenases, carbonic anhydrases, and phosphodiesterases). We expect that our method will be particularly successful when key protein/inhibitor interactions are conserved across all members of a protein family. Certain classes of proteases, such as matrix metalloproteases, appear to be particularly favorable candidates. Organizing 3D structural information by making a database of frameworks from pdb ligands in order to explore relationships between gene families and framework structures will be a focus of our future research.

Acknowledgment. We thank Drs. Paul Charifson, Mark Ledebor, Jeremy Green, John Thomson, and Mark Murcko for critically reading the manuscript and providing insightful comments.

Note Added after ASAP Posting. This manuscript was released ASAP on 7/27/2004 with incorrect numbers of frameworks indicated in the first paragraph of the Results section and with framework duplication and framework errors in Table 1. The correct version was posted on 8/13/2004.

References

- (1) Brooijmans, N.; Kuntz, I. D. Molecular recognition and docking algorithms. *Annu. Rev. Biophys. Biomol. Struct.* **2003**, *32*, 335–373.
- (2) Schneider, G.; Böhm, H.-J. Virtual screening and fast automated docking methods. *Drug Discovery Today* **2002**, *7*, 64–70.
- (3) Krumrine, J.; Raubacher, F.; Brooijmans, N.; Kuntz, I. Principles and methods of docking and ligand design. In *Structural Bioinformatics*; Bourne, P. E., Weissig, H., Eds.; Wiley-Liss, Inc.: Hoboken, NJ, 2003; pp 443–476.
- (4) Nissink, J. W. M.; Murray, C.; Hartshorn, M.; Verdonk, M. L.; Cole, J. C.; Taylor, R. A new test set for validating predictions of protein–ligand interaction. *Proteins* **2002**, *49*, 457–471.
- (5) Greenbaum, D. C.; Arnold, W. D.; Lu, F.; Hayrapetian, L.; Baruch, A.; Krumrine, J.; Toba, S.; Chehade, K.; Brömme, D.; Kuntz, I. D.; Bogoy, M. Small molecule affinity fingerprinting a tool for enzyme family subclassification, target identification and inhibitor design. *Chem. Biol.* **2002**, *9*, 1085–1094.
- (6) Baker, D.; Sali, A. Protein structure prediction and structural genomics. *Science* **2001**, *294*, 93–96.
- (7) Fiser, A.; Feig, M.; Brooks, C. L., 3rd; Sali, A. Evolution and physics in comparative protein structure modeling. *Acc. Chem. Res.* **2002**, *35*, 413–421.
- (8) Schwede, T.; Kopp, J.; Guex, N.; Peitsch, M. C. SWISS-MODEL: an automated protein homology-modeling server. *Nucleic Acids Res.* **2003**, *31*, 3381–3385.
- (9) Eswar, N.; John, B.; Mirkovic, N.; Fiser, A.; Ilyin, V. A.; Pieper, U.; Stuart, A. C.; Marti-Renom, M. A.; Madhusudhan, M. S.; Yerkovich, B.; Sali, A. Tools for comparative protein structure modeling and analysis. *Nucleic Acids Res.* **2003**, *31*, 3375–3380.
- (10) Bemis, G. W.; Murcko, M. A. The properties of known drugs. 1. Molecular frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.
- (11) Drevs, J.; Medinger, M.; Schmidt-Gersbach, C.; Weber, R.; Unger, C. Receptor tyrosine kinases: the main targets for new anticancer therapy. *Curr. Drug Targets* **2003**, *4*, 113–121.
- (12) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G. L.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The protein data bank. *Nucleic Acids Res.* **2000**, *28*, 235–242.
- (13) Shewchuk, L.; Hassell, A.; Wisely, B.; Rocque, W.; Holmes, W.; Veal, J.; Kuyper, L. F. Binding mode of the 4-anilinoquinazoline class of protein kinase inhibitor: X-ray crystallographic studies of 4-anilinoquinazolines bound to cyclin-dependent kinase 2 and P38 kinase. *J. Med. Chem.* **2000**, *43*, 133–138.
- (14) Pearson, W. R.; Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U.S.A.* **1988**, *85*, 2444–2448.
- (15) The kinase protein hinge is a segment connecting the two domains that form the ATP binding site. Jnk3 was chosen as the reference structure because it was an early kinase target for structure-based drug design at Vertex.
- (16) Xie, X.; Gu, Y.; Fox, T.; Coll, J. T.; Fleming, M. A.; Markland, W.; Caron, P. R.; Wilson, K. P.; Su, M. S. Crystal structure of JNK3: a kinase implicated in neuronal apoptosis. *Structure* **1998**, *6*, 983–991.
- (17) McLachlan, A. D. Rapid comparison of protein structures. *Acta Crystallogr.* **1982**, *A38*, 871–873.
- (18) Davies, T. G.; Bentley, J.; Arris, C. E.; Boyle, F. T.; Curtin, N. J.; Endicott, J. A.; Gibson, A. E.; Golding, B. T.; Griffin, R. J.; Hardcastle, I. R.; Jewsbury, P.; Johnson, L. N.; Mesguiche, V.; Newell, D. R.; Noble, M. E.; Tucker, J. A.; Wang, L.; Whitfield, H. J. Structure-based design of a potent purine-based cyclin-dependent kinase inhibitor. *Nat. Struct. Biol.* **2002**, *9*, 745–749.
- (19) DeLano, W. L. The PyMOL Molecular Graphics System (2002). <http://www.pymol.org>.
- (20) Stamos, J.; Sliwkowski, M. X.; Eigenbrot, C. Structure of the epidermal growth factor receptor kinase domain alone and in complex with a 4-anilinoquinazoline inhibitor. *J. Biol. Chem.* **2002**, *277*, 4265–4272.
- (21) 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*, 4339–4358.
- (22) Murtagh, F. A survey of recent advances in hierarchical clustering algorithms. *Comput. J.* **1983**, *26*, 354–359.

- (23) Murray, C. W.; Auton, T. R.; Eldridge, M. D. Empirical scoring functions. II. The testing of an empirical scoring function for the prediction of ligand–receptor binding affinities and the use of Bayesian regression to improve the quality of the model. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 503–519.
- (24) Eldridge, M. D.; Murray, C. W.; Auton, T. R.; Paolini, G. V.; Mee, R. P. Empirical scoring functions: I. The development of a fast empirical scoring function to estimate the binding affinity of ligands in receptor complexes. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 425–445.
- (25) Gibson, A. E.; Arris, C. E.; Benely, J.; Boyle, F. T.; Davis, N. J.; Curtin, T. G.; Endicott, J. A.; Golding, B. T.; Grant, S.; Griffin, R. J.; Jewsbury, P.; Johnson, L. N.; Mesguiche, V.; Newell, D. R.; Noble, M. E.; Tucker, J. A.; Whitfield, H. J. Probing the ATP ribose-binding domain of cyclin-dependent kinases 1 and 2 with O(6)-substituted guanine derivatives. *J. Med. Chem.* **2002**, *45*, 3381–3393.
- (26) Ikuta, M.; Kamata, K.; Fukasawa, K.; Honma, T.; Machida, T.; Hirai, H.; Suzuki-Takahashi, I.; Hayam, T.; Nishimura, S. Crystallographic approach to identification of cyclin-dependent kinase 4 (CDK4)-specific inhibitors by using CDK4 mimic CDK2 protein. *J. Biol. Chem.* **2001**, *276*, 27548–27554.
- (27) De Moliner, E.; Brown, N. R.; Johnson, L. N. Alternative binding modes of an inhibitor to two different kinases. *Eur. J. Biochem.* **2003**, *270*, 3174–3181.
- (28) Gray, N. S.; Wodicka, L.; Thunnissen, A. M.; Norman, T. C.; Kwon, S.; Espinoza, F. H.; Morgan, D. O.; Barnes, G.; LeClerc, S.; Meijer, L.; Kim, S. H.; Lockhart, D. J.; Schultz, P. G. Exploiting chemical libraries, structure and genomics in the search for kinase inhibitors. *Science* **1998**, *281*, 533–538.
- (29) Beattie, J. F.; Breault, G. A.; Ellston, R. P. A.; Green, S.; Jewsbury, P. J.; Midgley, C. J.; Naven, R. T.; Minshull, C. A.; Pauptit, R. A.; Tucker, J. A.; Pease, J. E. Cyclin-dependent kinase 4 inhibitors as a treatment for cancer. Part 1: identification and optimisation of substituted 4,6-bis anilino pyrimidines. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2955–2960.
- (30) Dreyer, M. K.; Borcharding, D. R.; Dumont, J. A.; Peet, N. P.; Tsay, J. T.; Wright, P. S.; Bitonti, A. J.; Shen, J.; Kim, S.-H. Crystal structure of human cyclin-dependent kinase 2 in complex with adenine-derived inhibitor H717. *J. Med. Chem.* **2000**, *44*, 524–530.
- (31) Nagar, B.; Bornmann, W.; Pellicena, P.; Schindler, T.; Veach, D.; Miller, W. T.; Clarkson, B.; Kuriyan, J. Crystal structures of the kinase domain of C-Abl in complex with the small molecule inhibitors Pd173955 and Imatinib (Sti-571). *Cancer Res.* **2002**, *62*, 4236–4243.
- (32) Klutchko, S. R.; Hamby, J. M.; Boschelli, D. H.; Wu, Z.; Kraker, A. J.; Amar, A. M.; Hartl, B. G.; Shen, C.; Klohs, W. D.; Steinkampf, R. W.; Driscoll, D. L.; Nelson, J. M.; Elliott, W. L.; Roberts, B. J.; Stoner, C. L.; Vincent, P. W.; Dykes, D. J.; Panek, R. L.; Lu, G. H.; Major, T. C.; Dahring, T. K.; Hallak, H.; Bradford, L. A.; Showalter, H. D.; Doherty, A. M. 2-Substituted Aminopyrido[2,3-*d*]pyrimidin-7(8*H*)-ones. Structure–activity relationships against selected tyrosine kinases and in vitro and in vivo anticancer activity. *J. Med. Chem.* **1998**, *41*, 3276–3292.
- (33) Mohammadi, M.; Froum, S.; Hamby, J. M.; Schroder, M. C.; Panek, R. L.; Lu, G. H.; Eliseenkova, A. V.; Green, D.; Schlessinger, J.; Hubbard, S. R. Crystal structure of an angiogenesis inhibitor bound to the Fgf receptor tyrosine kinase domain. *EMBO J.* **1998**, *17*, 5896–5904.
- (34) Bramson, H. N.; Corona, J.; Davis, S. T.; Dickerson, S. H.; Edelstein, M.; Frye, S. V.; Gampe, R. T.; Hassell, A. H.; Shewchuk, L. M.; 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*, 4339–4358.
- (35) De Moliner, E. D.; Moro, S.; Sarno, S.; Zagotto, G.; Zanotti, G.; Pinna, L.A.; Battistutta, R. Inhibition of protein kinase CK2 by anthraquinone-related compounds. *J. Biol. Chem.* **2003**, *278*, 1831–1836.

JM0499054