Use of Catalyst Pharmacophore Models for Screening of Large Combinatorial Libraries

Evan A. Hecker, Chaya Duraiswami, Tariq A. Andrea, and David J. Diller*

Department of Molecular Modeling, Pharmacopeia, Inc., CN5350, Princeton, New Jersey 08543-5350

Received March 1, 2002

Using a data set comprised of literature compounds and structure−activity data for cyclin dependent kinase 2, several pharmacophore hypotheses were generated using Catalyst and evaluated using several criteria. The two best were used in retrospective searches of 10 three-dimensional databases containing over 1 000 000 proprietary compounds. The results were then analyzed for the efficiency with which the hypotheses performed in the areas of compound prioritization, library prioritization, and library design. First as a test of their compound prioritization capabilities, the pharmacophore models were used to search combinatorial libraries that were known to contain CDK active compounds to see if the pharmacophore models could selectively choose the active compounds over the inactive compounds. Second as a test of their utility in library design again the pharmacophore models were used to search the active combinatorial libraries to see if the key synthons were over represented in the hits from the pharmacophore searches. Finally as a test of their ability to prioritize combinatorial libraries, several inactive libraries were searched in addition to the active libraries in order to see if the active libraries produced significantly more hits than the inactive libraries. For this study the pharmacophore models showed potential in all three areas. For compound prioritization, one of the models selected active compounds at a rate nearly 11 times that of random compound selection though in other cases models missed the active compounds entirely. For library design, most of the key fragments were over represented in the hits from at least one of the searches though again some key fragments were missed. Finally, for library prioritization, the two active libraries both produced a significant number of hits with both pharmacophore models, whereas none of the eight inactive libraries produced a significant number of hits for both models.

INTRODUCTION

Over the past 30 years, computational methods and molecular modeling programs have been increasingly used for the interpretation of structure activity data and to guide further compound synthesis during lead optimization. One of the most common such methods is pharmacophore modeling embodied in programs such as Catalyst.[1] A pharmacophore model is defined as the three-dimensional arrangement of the structural and physicochemical features that are relevant to biological activity.[2] With the growing use of pharmacophore modeling, it is important to critically assess its validity for its common applications and examine whether there are additional problems that could be tackled with these methods—particularly those problems associated with data mining large combinatorial libraries.

Pharmacophore modeling is a versatile tool to aid in the discovery and development of new lead compounds.[3] It has been shown that as an intuitive device, a pharmacophore model can guide chemists to include functional groups and chemical features in the compounds synthesized during lead optimization thereby improving the chance of finding potent compounds.[4,5] Pharmacophore models have also been used to estimate binding affinities to determine which compounds offer the greatest chance to be active, and upon synthesis of these compounds the pharmacophore models have been shown in some cases to be predictive.[6,7] Pharmacophore models have been further applied to the discovery of new lead compounds through the searching of large compound databases,[8−10] i.e., the prioritization of compound screening. As an example of a typical such use, Nicklaus et al.[10] used a known inhibitor of HIV-1 integrase to generate a pharmacophore model, search a proprietary database, and prioritize 267 possible leads. Sixty of those were ultimately tested, and 19 were found to inhibit the enzyme. Finally, pharmacophore models have been used in combinatorial library design to both ensure the diversity of the compounds in the library[11,12] and to bias the library toward compounds containing a known "active" pharmacophore.

The utility of pharmacophore modeling to prioritize compounds via database searches and the utility of pharmacophore models to improve library design are important concepts, but pharmacophore models have the potential for uses in other areas as well. Beyond library design and compound prioritization, pharmacophore models could also be used as library prioritization tools, i.e., to select combinatorial libraries that have a higher frequency of the active pharmacophore and thus a better chance of containing potent compounds. It is these three uses, compound prioritization, library design, and library prioritization, that we investigate in this study.

Programs such as Catalyst[1] can be used to transform data sets of active and inactive compounds into three-dimensional pharmacophore models. Catalyst generates hypotheses (pharmacophore models) by correlating biological activity and chemical structure.[9] The resulting pharmacophore models are

* Corresponding author phone: (609)452-3783; fax: (732)422-0156; e-mail: ddiller@pharmacop.com.

10.1021/ci020368a CCC: $22.00
Published on Web 08/15/2002

composed of three-dimensional representations of the functional groups and chemical features that are necessary to separate active from inactive compounds. The process for the creation of hypotheses is well documented.[9,13−17]

Using a data set comprised of literature compounds and structure−activity data, several pharmacophore hypotheses were generated and evaluated. The two best were utilized in searches of three-dimensional databases containing over 1 000 000 proprietary compounds. The results were then analyzed for the efficiency with which the hypotheses performed in compound prioritization, library prioritization, and library design. This analysis suggests that pharmacophore models are indeed useful in these three areas and that using multiple pharmacophore models can improve results.

Cyclin-dependent kinases, CDK, are a class of protein kinases, which play a key role in cell cycle regulation and recently have become the target of pharmaceutical focus.[18] Inhibitors of CDK2, for example, appear to be potential therapeutic agents for controlling aberrant cell propagation in a range of diseases, including cancer and Alzheimer's disease.[18] The crystal structure and mechanism of CDK2 activation are known.[19,20] In addition, CDK2 structure−activity data are available in the literature including 2,6,9-trisubstituted purines.[21] Finally, in house high throughput screening results of over 4 000 000 compounds are available for validation. Thus CDK2 makes a very interesting and nearly ideal validation case. As a result, CDK2 was chosen as the system for this retrospective study.

METHODS

The training set of CDK2 inhibitors used in this study was taken from several literature sources.[21−25] The data set consists of 35 compounds, 34 of which are purine derivatives, see Figure 1. The CDK2 IC_{50} values vary over 3 orders of magnitude, from 6 nM−7 μM, whereas typically a range in potency of 4 orders of magnitude is preferred for deriving a Catalyst hypothesis. Also, only two of these compounds have low nanomolar activity: compounds 1 and 2 have IC50 values of 6 nM and 9 nM, respectively. Finally, for the purposes of this study we merged high throughput screening data for CDK2 and closely related homologues. Within the CDK family of kinases there are several closely related homologues including CDK1, CDK2, CDK3 and CDK5. These four CDKs are extremely similar in amino acid sequence and nearly identical in their binding sites, see Figure 2. In fact, the only significant differences in the binding site are in the residues of the hinge region, Figure 2. The side chains of these residues all point out of the binding site and therefore are likely to have minimal impact on potency. Additionally, IC50s for active compounds across these four kinases rarely vary more (∼factor of 10) than the IC50s for the same compound and same CDK with different cyclins.[22,23,26−29] In practice one never has a perfect data set at the beginning of a discovery project. Thus this case makes a realistic case for a validation study.

To create the pharmacophore models used in this study, several Catalyst runs were performed varying only in the number of features set as a minimum generating a total of 54 hypotheses. These hypotheses contained a minimum of three and a maximum of five features with a varying number of hydrophobic regions, hydrogen bond acceptors, hydrogen bond donors, aromatic rings, and positive ionizable groups. The hypotheses were rated with two methods. A regression method was performed first on all hypotheses, followed by a compare/fit analysis on the six best hypotheses. The regression analysis rates each hypothesis by estimating the activity for each compound in the data set and then comparing the estimated activity to the experimentally determined binding constants. This method is commonly utilized in other Catalyst studies.[6]

Six hypotheses were chosen from the results of the regression analysis—two 3-feature models, two 4-feature models, and two 5-feature models. These models were then evaluated using a compare/fit analysis as follows: the two most active compounds, molecules 1 and 2, were mapped to the six hypotheses using the best flexible fit. The estimated activity and conformational strain were recorded for all mappings within 1 order of magnitude of the estimated activity of the best fit. The two hypotheses that mapped to these two compounds and demonstrated accurate estimated activities with conformers of low strain were considered the best and were chosen for the database searches (Figure 3). Using these selection criteria, a 4-feature (Figure 3a) and a 5-feature (Figure 3b) hypothesis were chosen. These two hypotheses also showed the best correlation with the data: the correlation coefficients (r^2) for these two hypotheses are 0.74 for the 4-feature model and 0.68 for the 5-feature model.

Preliminary database searches showed that the 4-feature hypothesis was too permissive, generating a number of hits close to the total in the database. In addition, the compounds selected by the 4-feature hypothesis were often clearly too large to fit in and ultimately bind in the CDK2 binding site. As a result, the 4-feature hypothesis was merged with a shape query based on the best mapping of the most active molecule, molecule 1, to the hypothesis. The hypotheses will henceforth be referred to as 4-feature-shape and 5-feature, Figure 3.

The two hypotheses, 4-feature-shape and 5-feature, were used to search through over 1 000 000 proprietary compounds encompassing ten ECLIPS combinatorial libraries.[30] The libraries were chosen because they represent a large range of structural classes. Five of these libraries are considered "kinase" libraries due to their history of producing inhibitors for a variety of kinases. The other five were selected because they contain roughly the same number of compounds (∼100 000) as the five kinase libraries. Of the five kinase libraries, two, LIBa1 and LIBa2, have produced active compounds for at least one of the four CDKs described above, 39 total actives from LIBa1 and 17 total actives from LIBa2, from the high throughput screens. The other three kinase libraries, LIBk1, LIBk2, and LIBk3, have not produced any CDK actives. None of the five remaining libraries, LIBi1, LIBi2, LIBi3, LIBi4, and LIBi5, produced any CDK active compounds.

RESULTS AND DISCUSSION

Hypothesis generation and evaluation were performed as described above. The two best hypotheses were chosen for subsequent study. The 4-feature-shape hypothesis consists of a hydrogen bond acceptor, an aromatic ring, and two hydrophobic groups, Figure 3a. The 5-feature hypothesis consists of the same chemical features with the addition of a positive ionizable group, Figure 3b. It does not, however,

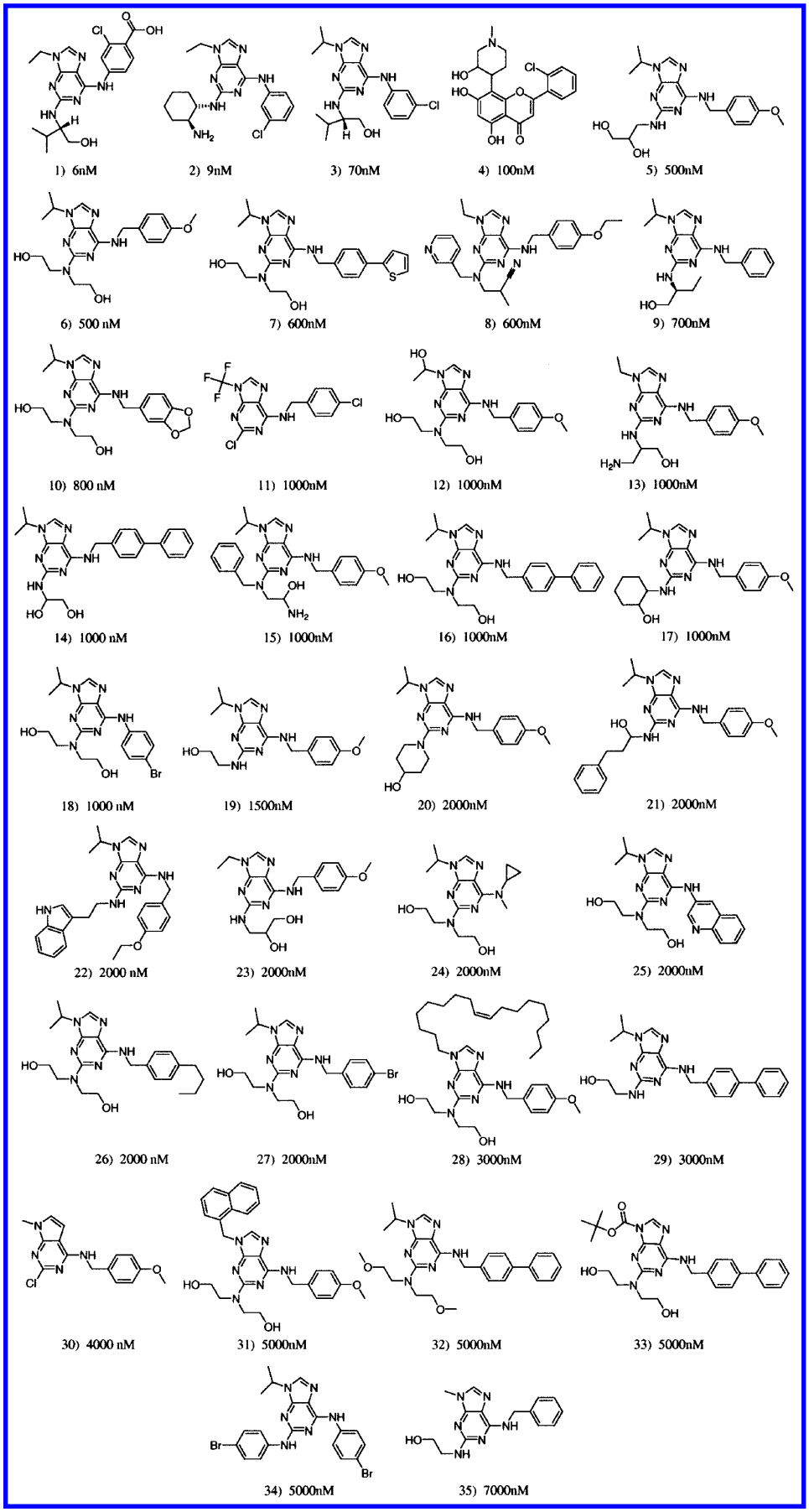

Figure 1. The compounds in the training set used to build the pharmacophore models.

contain a shape feature. In addition, the features are spatially arranged differently between the two hypotheses. The only compound that has a positive ionizable group is compound 2. Thus one might wonder how the other active compounds

```
CDK2          1 MENFQKVEKIGEGTYGVVYKARNKLTGEVVALKKIRLDTETEGVPSTAIREISLLKELNH  60
CDK3            MDMFQKVEKIGEGTYGVVYKAKNRETGQLVALKKIRLDLEMEGVPSTAIREISLLKELKH
CDK1            MEDYTKIEKIGEGTYGVVYKGRHKTTGQVVAMKKIRLESEEEGVPSTAIREISLLKELRH
CDK5            MQKYEKLEKIGEGTYGTVFKAKNRETHEIVALKRVRLDDDDEGVPSSALREICLLKELKH
                *:  : *:*********.*:*.::: * ::**:*::**: : *****:*:***.*****.*
                                  Hinge Region
CDK2         61 PNIVKLLDVIHTENKLYLVFEFLHQDLKKFMDASA-LTGIPLPLIKSYLFQLLQGLAFCH 120
CDK3            PNIVRLLDVVHNERKLYLVFEFLSQDLKKYMDSTP-GSELPLHLIKSYLFQLLQGVSFCH
CDK1            PNIVSLQDVLMQDSRLYLIFEFLSMDLKKYLDSIPPGQYMDSSLVKSYLYQILQGIVFCH
CDK5            KNIVRLHDVLHSDKKLTLVFEFCDQDLKKYFDSCN--GDLDPEIVKSFLFQLLKGLGFCH
                 *** * **:  : :* *:***   ****::*:      :   ::**:*:*:*:*: ***

CDK2        121 SHRVLHRDLKPQNLLINTEGAIKLADFGLARAFGVPVRTYTHEVVTLWYRAPEILLGCKY 180
CDK3            SHRVIHRDLKPQNLLINELGAIKLADFGLARAFGVPLRTYTHEVVTLWYRAPEILLGSKF
CDK1            SRRVLHRDLKPQNLLIDDKGTIKLADFGLARAFGIPIRVYTHEVVTLWYRSPEVLLGSAR
CDK5            SRNVLHRDLKPQNLLINRNGELKLADFGLARAFGIPVRCYSAEVVTLWYRPPDVLFGAKL
                *:.*:***********:  * :************:*:* *: ********.*::*:*.

CDK2        181 YSTAVDIWSLGCIFAEMVTR-RALFPGDSEIDQLFRIFRTLGTPDEVVWPGVTSMPDYKP 240
CDK3            YTTAVDIWSIGCIFAEMVTR-KALFPGDSEIDQLFRIFRMLGTPSEDTWPGVTQLPDYKG
CDK1            YSTPVDIWSIGTIFAELATK-KPLFHGDSEIDQLFRIFRALGTPNNEVWPEVESLQDYKN
CDK5            YSTSIDMWSAGCIFAELANAGRPLFPGNDVDDQLKRIFRLLGTPTEEQWPSMTKLPDYKP
                *:*.:*:** * ****:..  :.** *:.  *** **** **** :  ** : .: ***

CDK2        241 SFPKWARQ-DFSKVVPPLDEDGRSLLSQMLHYDPNKRISAKAALAHPFFQDVTKPVPHLR 300
CDK3            SFPKWTRK-GLEEIVPNLEPEGRDLLMQLLQYDPSQRITAKTALAHPYFSSP-EPSPAAR
CDK1            TFPKWKPG-SLASHVKNLDENGLDLLSKMLIYDPAKRISGKMALNHPYFNDLDNQIKKM-
CDK5            -YPMYPATTSLVNVVPKLNATGRDLLQNLLKCNPVQRISAEEALQHPYFSDFCPP-----
                 :* :    .: . *  *:  * .** ::*  :* :**:.: ** **:*..
```

Figure 2. A multiple sequence alignment of the cyclin dependent kinases, CDK1 CDK2, CDK3, and CDK5. The residues in the binding site are shown in bold.

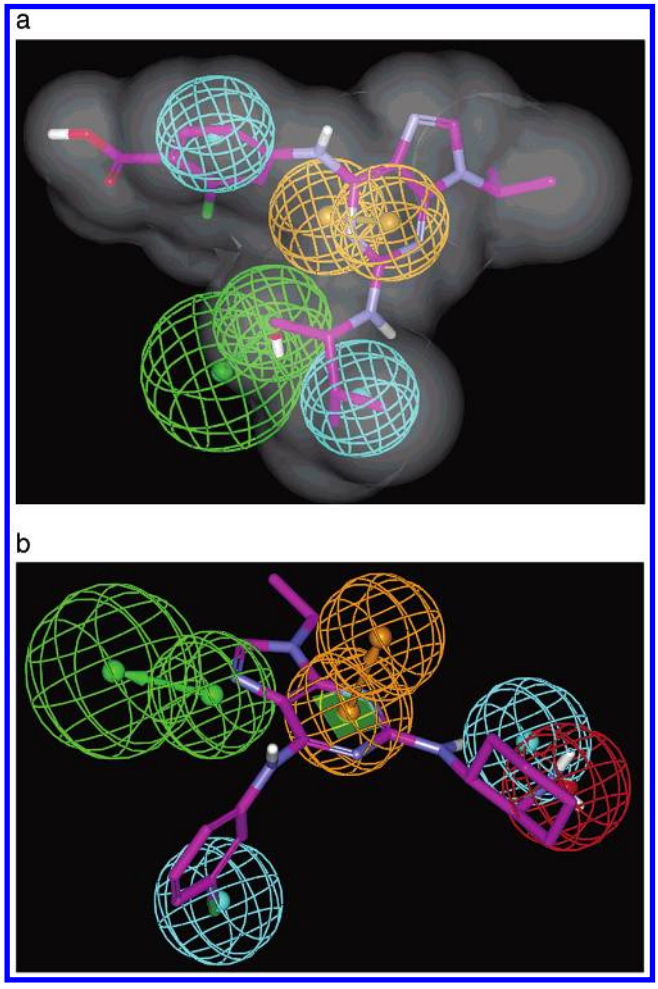

Figure 3. The hypotheses. (a) Molecule 1 mapped to the 4-feature-shape pharmacophore model. The light blue feature is a hydrophobic feature. The orange feature is an aromatic feature. The green feature is a hydrogen bond donor. (b) Molecule 2 mapped to the 5-feature pharmacophore model. The red feature is a positive ionizable group. The color coding for the remaining features is the same as for (a).

could map to this model with a high predicted potency. While compounds 1 and 3 cannot map anything to the positive ionizable feature, they do map the remaining features more accurately resulting in comparable activity predictions.

Table 1. Enrichment Results within the Two Active Libraries[a]

		4-feature shape				5-feature			
library	HTS actives	hit %	#act. hits	act. hit %	enrich.	hit %	#act. hits	act. hit %	enrich.
LIBa1	39	9.2	4	10.3	1.1	3.8	16	41.0	10.8
LIBa2	17	8.5	4	23.5	2.8	10.2	0	0.0	0.0

[a] The HTS actives column is the number of actives found in the given library via high throughput screening. The hit % column is the percentage of compounds from the library that fit the given hypothesis. The #act. Hits column is the number of the HTS actives from the given library that fit the given hypothesis. The act. hit% column is the percentage of the HTS actives that fit the give hypothesis. The enrich. column is the enrichment as calculated by eq 1.

The 4-feature-shape and 5-feature hypotheses were used as queries in the three-dimensional database searches of the 10 proprietary combinatorial libraries amounting to over 1 000 000 compounds. The results were then analyzed and evaluated on three grounds: compound prioritization, library design, and library prioritization. The overall results of all of the library searches are described below and briefly summarized in Tables 1 and 2.

A Comparison to the X-ray Structure. The structure 1ckp[31] was used to compare the hypotheses to the CDK2 binding site. This structure was cocrystallized with a compound similar to many of the purines in the training set. The optimal rotation and translation required to map the purine ring of compound 1 as fit to the hypothesis to the purine ring of the inhibitor in the X-ray structure was used to bring each hypothesis as a rigid body into the reference frame of the binding site. A summary of the mappings of the models to the X-ray structure is shown in Figure 4.

The aromatic ring feature (orange) in both models is stacked directly against the side chain of Leu134. This is a common kinase/inhibitor interaction and is likely important. The hydrogen bond acceptor feature of the 5-feature model is positioned to interact with the NH of the hinge region

Table 2. Results of 10 Library Searches with Two Hypotheses[a]

library	HTS actives	# cpds	4-feature/shape # hits	4-feature/shape hit %	5-feature # hits	5-feature hit %	geometric mean
LIBa1	39	117180	10787	9.2	4478	3.8	5.9
LIBa2	17	106596	9053	8.5	10832	10.2	9.3
LIBk1		181629	923	0.5	13079	7.2	1.9
LIBk2		11718	639	5.4	94	0.8	2.1
LIBk3		60543	2489	4.1	2179	3.6	3.8
LIBi1		100800	553	0.6	52	0.05	0.2
LIBi2		107100	212	0.2	105	0.1	0.1
LIBi3		96100	33	0.03	49	0.05	0.04
LIBi4		133875	621	0.5	32227	24.1	3.3
LIBi5		139725	678	0.5	20065	14.4	2.6
total		1055266	25988	2.5	83160	7.9	

[a] The HTS actives column is the number of active compounds found from the high throughput screen in the given library. The geometric mean is the geometric mean of the hit percentages from the two hypotheses.

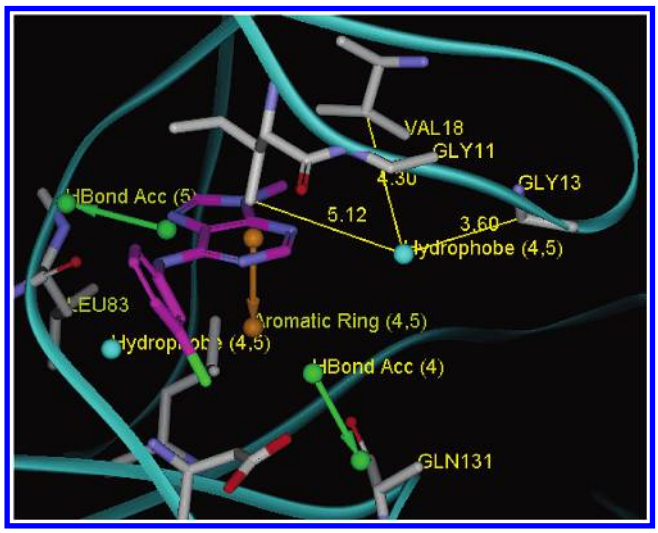

Figure 4. The hypotheses mapped to a CDK2 X-ray structure. The carbon atoms of the CDK2 residues that are shown are shown in gray. The ligand (purple) shown is that cocrystallized with CDK2. The features are labeled according to their type, and the model to which they belong is given in parentheses. When two features of the same type but from different models were in the same location only one is shown. The only feature not shown is the positive ionizable feature of the 5-feature model. This is found between the second hydrophobic feature the tail of the hydrogen bond acceptor from the 4-feature model. The aromatic feature is shown in orange. The hydrogen bond acceptor features are shown in green. The hydrophobic feature is shown in light blue.

residue of Leu83. This is again a commonly observed interaction in kinase/inhibitor complexes and thus likely important. Because the purine rings were used to map the hypotheses to the binding site and these features were on the purine ring system they were likely to be mapped to significant interaction sites on the protein.

The hydrophobic feature (light blue) mapped to the Cl-phenyl in both models is mapped roughly to the secondary hydrophobic pocket where the Cl-phenyl of the bound inhibitor also sits. The second hydrophobic feature of both models is mapped close to a hydrophobic pocket formed by the active site loop. This pocket consists of residues Ile10, Gly11, Gly13, and Val18. The positioning of these two hydrophobic features is consistent with the binding modes described for compounds 1 and 2.[22,23]

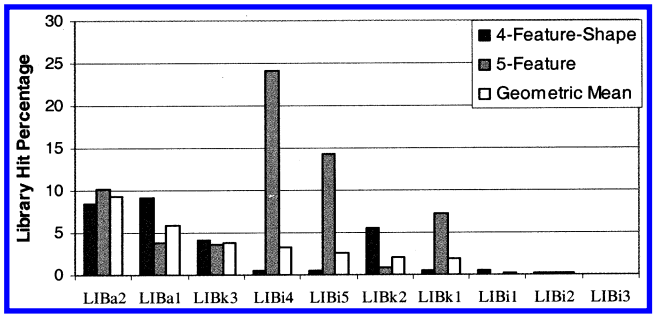

Figure 5. The percentage of hits from each library with each hypothesis. The libraries are ordered using the geometric mean of the results with the two hypotheses.

Finally, the hydrogen bond acceptor feature (green) of the 4-feature-shape model mapped between the side chain of Asp 86 and the carbonyl oxygen of Gln 131. The positive ionizable feature (red) of the 5-feature model is positioned between the hydrogen bond acceptor of the 4-feature model and the second hydrophobic feature and is still roughly in the vicinity of the carbonyl oxygen of Gln131. While the positioning of these features is consistent with the reported binding modes of compounds 1 and 2,[22,23] the most general feature for this area is a hydrogen bond donor. The hydrogen bond acceptor of the 4-feature model was likely found because the molecules used in the training set all had alcohols at this position and thus could act as either a hydrogen bond acceptor or a hydrogen bond donor. As a result the structure activity data alone was insufficient to determine the hydrogen bonding preference in this area.

The only obvious interaction missed by both models is the hydrogen bond between the NH of the Cl-phenyl-amino and the carbonyl oxygen of Leu86. This might have been missed because nearly all the inhibitors have this NH, and thus there is no way to determine its relevance. Ultimately, based on the information available in the crystal structure, the two models could be merged into one that would consist of the two hydrophobic features, the aromatic feature, the hydrogen bond acceptor of the 5-feature model, and a hydrogen bond donor to replace both the hydrogen bond acceptor of the 4-feature model and the positive ionizable feature of the 5-feature model. Because an X-ray structure is not typically available early in a discovery project the information extracted from the X-ray structure was not used in the remainder of the study, and the two models were used as originally built.

Compound Prioritization. As described above, several relevant active compounds were discovered in libraries LIBa1 and LIBa2 from high throughput screens. There were 39 active compounds discovered from a total of 117180 compounds in LIBa1 and 17 unique active compounds discovered from a total of 106596 compounds in LIBa2. These active compounds and libraries were used to validate the hypotheses' ability to preferentially select active compounds over inactive compounds. The measurement of the quality of compound prioritization used is the enrichment factor.[14] Loosely speaking, the enrichment of a database search is the ratio of the number of active compounds in the selected compounds over the number of active compounds that would have been expected to have been found had the same number of compounds been selected at random.

Mathematically,

$$\text{enrichment} = \frac{\frac{a}{n}}{\frac{A}{N}} \quad (1)$$

where N is the number of compounds in the library, A is the number of active compounds in the library, n is the number of selected hits, i.e., the compounds deemed to match the pharmacophore model, and a is the number of active compounds in the selected compounds (active hits).

The 4-feature-shape hypothesis exhibited mild enrichment for both libraries. Of the 117 180 compounds in LIBa1, 10 787 matched this hypothesis. Of these 10 787 hits, 4 were active. Thus by eq 1 the enrichment is 1.1. Thus for LIBa1 the 4-feature-shape hypothesis performed no better than random compound selection. Of the 106 596 compounds in LIBa2, 9053 matched this hypothesis. Of these 9053 hits, again 4 were active. But because there were only 17 actives in LIBa2, the resulting enrichment is 2.8. While the 4-feature-shape hypothesis showed mild enrichment for both LIBa1 and LIBa2, the 5-feature hypothesis was more disparate between the two libraries. For LIBa1, 4478 compounds matched the 5-feature hypothesis. Of these 4478 compounds, 16 were active resulting in an enrichment of 10.8. For LIBa2, 10 832 compounds matched the 5-feature hypothesis. None of these 10 832 compounds, however, was active.

A possible rationalization for these enrichment results is that the 5-feature hypothesis is too biased by a single compound in the training set. Compound 2 is the only potent compound with a positive ionizable center and thus the positive ionizable feature of the 5-feature hypothesis was chosen largely due to this compound. As most of the HTS actives found in LIBa1 contain a positive ionizable center this model performed well on this library. But the majority of the HTS actives from LIBa2 do not contain the positive ionizable center and thus were not found by the 5-feature hypothesis. While the 4-feature-shape hypothesis did not perform as well on LIBa1, it does appear to be more general.

Library Prioritization. Instead of looking for a specific lead, a hypothesis could be used to narrow the search to a library, which can then be searched using high-throughput screening techniques for a lead to be optimized. The same two hypotheses as described above were used for this study. The searches were performed on the ten combinatorial libraries, totaling over 1 000 000 compounds. The results of the 20 database searches are shown in Table 2. To rank libraries, the percentage of the compounds in the library that match the pharmacophore was used.

In only three of the libraries did more than three percent of the compounds match both pharmacophore models: LIBa1, LIBa2, and LIBk3. Of these three libraries, the two libraries hit the most by the two hypotheses were LIBa1 and LIBa2, the two active libraries. This is encouraging as it shows the hypotheses were able to correctly prioritize the libraries. It is also surprising that while the 5-feature hypothesis showed no enrichment for LIBa2 it found more hits in LIBa2 than it did in LIBa1 where it showed significant enrichment. While it was unable to locate active compounds within the LIBa2 it still indicated, by the overall hit rate, that the library was more likely to contain actives than the other libraries.

Table 3. Library Design Results for LIBa1[a]

synthon	no. of HTS actives	HTS active occurrence	5-feature occurrence	4-feature occurrence
R1a12	3	4.7	0.6	2.0 (6)
R1a1	2	3.2	0.9	1.6 (10)
R1a15	2	3.2	1.1 (10)	1.8 (8)
R1a19	2	3.2	0.3	0.7
R1a30	2	3.2	0.5	0.2
R1a37	2	3.2	0.5	0.6
R1a54	2	3.2	0.8	0.9
R1a62	2	3.2	0.9	1.2 (16)
R2a30	17	13.2	3.0 (1)	0.70
R2a15	3	2.3	0.5	3.4 (1)
R2a8	2	1.6	1.1 (12)	0.8
R2a18	2	1.6	0.2	0.3
R2a20	2	1.6	0.1	0.1
R2a27	2	1.6	0.2	0.4
R3a16	19	28.5	8.4 (1)	1.0
R3a26	21	31.5	0.5	0.3

[a] LIBa1 was synthesized in three steps. There are 63 synthons for R1, 31 synthons for R2, and 60 synthons for R3. The first column is the synthon name. The second column is the number of the high throughput actives containing the given synthon. The three occurrence columns are the occurrence of the synthon in the list divided by the expected occurrence of the synthon. A number in parentheses is the rank of this synthon for the given hypothesis.

A few of the inactive libraries did have a significant number of compounds match either of the hypotheses. For example, for the searches of LIBi4 and LIBi5, many hits were found with the 5-feature hypothesis, while 4-feature-shape hypothesis showed few hits. The 5-feature hypothesis found many compounds with similar chemical features but neglected the size of the molecule. In fact, the library contained rather large compounds, which would likely be unable to fit in the binding site. The 4-feature-shape hypothesis did not hit these compounds likely because the shape filter negated them as hits. Thus, the combination of the two hypotheses greatly reduced false positives and as a result, significantly improved the results. Indeed, merging a shape feature with 5-feature hypothesis might have produced the same effect.

Library Design. Finally, the results of the database searches were examined for their potential use in library design. In this aspect of the study, compound fragments were the focus rather than whole compounds or entire libraries. The hits from the two libraries with active compounds, LIBa1 and LIBa2, were analyzed with regard to the frequency of the occurrence of each synthon in the hits compared to the frequency of their occurrence in the high throughput actives. To assign a quantitative score to a synthon from a hit list (either from one of the hypotheses or from the high throughput actives) we use the occurrence which is defined as the number of times the synthon appears in the hit list divided by the mean number of times a synthon at the given position appeared in the hit list. An occurrence greater than 1 means the synthon was over represented in the hit list, whereas an occurrence less than 1 means the synthon was under represented in the hit list. The overall library design results are summarized in Tables 3 and 4.

LIBa1 was synthesized in three steps (see Table 3). At R1 there are 63 synthons, at R2 there are 31 synthons, and

Table 4. Library Design Results for LIBa2[a]

synthon	no. of HTS actives	HTS active occurrence	5-feature occurrence	4-feature occurrence
R1a42	2	7.4	1.2 (13)	3.3 (2)
R1a45	2	7.4	0.9	2.8 (4)
R1a8	1	3.7	1.2 (11)	1.6 (11)
R1a10	1	3.7	1.2 (9)	0.8
R1a12	1	3.7	0.7	0.1
R1a22	1	3.7	0.9	0.4
R1a35	1	3.7	0.8	1.6 (13)
R1a44	1	3.7	0.8	1.1
R1a48	1	3.7	1.1	0.9
R1a50	1	3.7	1.4 (5)	3.5 (1)
R1a52	1	3.7	1.0	0.5
R1a54	1	3.7	1.0	2.6 (5)
R1a57	1	3.7	1.6 (2)	3.3 (3)
R1a58	1	3.7	1.3 (6)	1.2
R1a62	1	3.7	1.0	1.0
R2a10	10	21.2	2.7 (3)	1.3 (12)
R2a9	3	6.4	2.7 (2)	1.0 (17)
R2a36	2	4.2	2.0 (7)	1.1 (14)
R2a15	2	4.2	2.8 (1)	1.0
R3a3	17	47	0.1	0.8

[a] LIBa2 was synthesized in three steps. There are 63 synthons for R1, 36 synthons for R2, and 47 synthons for R3. The columns are defined as in Table 2.

at R3 there are 60 synthons. Of the active synthons at R1, only one appears in more than two decoded compounds, R1a12. In the hit list from the search with the 4-feature-shape hypothesis, the R1a12 synthon has occurrence of 2.0, which is the sixth highest occurrence of all the R1 synthons in this hit list. Several of the remaining active synthons at R1 have occurrences above 1 for the 4-feature-shape hypothesis. Thus the 4-feature-shape hypothesis appears to preferentially select "active" synthons at R1. The 5-feature hypothesis, however, nearly missed the active synthons at R1 altogether. At the R2 position, the highest ranked synthon by the 5-feature hypothesis, R2a30, is the key synthon at this position. R2a30 appears in 17 of the 39 decodes. Also, at the R2 position, the highest ranked synthon by the 4-feature-shape hypothesis, R2a13, appeared in 3 decodes. Thus between the two hypotheses the top two synthons at R2 were clearly selected. At the R3 position the top ranked synthon by the 5-feature hypothesis, R3a16, is one of only two synthons appearing in the high throughput actives. The other active synthon, R3a26, was not preferentially selected by either of the hypotheses. Overall for LIBa1, the pharmacophore models again showed some correlation to the screening data with respect to selection of preferred synthons and when combined there were fewer missed fragments. One key fragment, R3a26, was, however, missed by both hypotheses.

The same analysis was performed on the results of the searches of LIBa2. A comparable correlation was observed. LIBa2 was again synthesized in three steps with 63 synthons at R1, 36 synthons at R2, and 47 synthons at R3. At R1 there were no clear preferred fragments in the decoded compounds (see Table 4). Indeed, 15 of the synthons at R1 appeared in the decoded compounds with only two of them appearing twice. Of these 15 synthons, several were highly ranked by at least one of the hypotheses. The R2 position appears to show the best correlation with the data. In this case there are four active synthons, all of which have an occurrence of 1.0 or greater with both hypotheses. In particular, the 5-feature hypothesis ranked these synthons as 1, 2, 3, and 7. For the R3 position only one synthon, R3a3, appears in the decoded compounds. This synthon is not preferentially selected by either hypothesis. Thus as with LIBa1 at the synthon level, there appears to be some correlation between the pharmacophore hits and the decoded compounds but some fragment information was missed. It is possible that an additional pharmacophore model would be sensitive to such omissions.

CONCLUSION

Using a data set comprised of CDK2 inhibitors taken from the literature, several hypotheses were generated using the hypogen module of Catalyst. The hypotheses were then evaluated based solely on this literature data with the two best being selected for the validation experiments. A 4-feature-shape hypothesis and a 5-feature hypothesis (see Figure 3) were used in the three-dimensional database searches of 10 proprietary combinatorial libraries, totaling more than 1 000 000 compounds. The results were analyzed for three potential uses: compound prioritization, library prioritization, and library design. Neither hypothesis exhibited significant enrichment in both active libraries. The 4-feature-shape hypothesis did show enrichment for LIBa2, while the 5-feature hypothesis showed enrichment in LIBa1. Thus by using both hypotheses active compounds from both libraries could have been identified.

Both the hypotheses found a significant number of hits from both active libraries, LIBa1 and LIBa2, in comparison to the number of hits found in the inactive libraries. This finding indicates library prioritization employing pharmacophore models is practical. Again, if used individually, there would have been some false positives (particularly with the 5-feature hypothesis) but when used together the two active libraries clearly stood out.

The hits from the database searches were also examined to evaluate the potential of pharmacophore models for use in combinatorial library design. The pharmacophore models were seen to preferentially select the fragments that were over represented in the active compounds. In this case, using only one hypothesis would have resulted in missing a number of important synthons. By using both pharmacophore models, far fewer of the key fragments were missed. Even with both pharmacophore models, however, two of the key fragments were missed. Indeed it is possible that additional pharmacophore models would have found these last two key synthons.

REFERENCES AND NOTES

(1) Catalyst v4.5. Molecular Simulations, I; San Diego, CA.
(2) Ghose, A. K.; Wendoloski, J. J. Pharmacophore Modelling: Methods, Experimental Verification and Application. *Perspect. Drug Discovery Des.* **1998**, 253−271.
(3) Charifson, P. S. *Practical Application of Computer-Aided Drug Design*; Marcel Dekker: New York, 1997.
(4) Traxler, P.; Green, J.; Mett, H.; Sequin, U.; Furet, P. Use of a Pharmacophore Model for the Design of EGFR Tyrosine Kinase Inhibitor: Isoflavones and 3-Phenyl-4(1H)-quinolones. *J. Med. Chem.* **1999**, *42*, 1018−1026.
(5) Ward, J. S.; Merritt, L.; Calligaro, D. O.; Bymaster, F. P.; Shannon, H. E.; Mitch, C. H.; Whitesitt, C.; Brunsting, D.; Sheardown, M. J.; Olesen, P. H.; Swedberg, M. D. B.; Jeppesen, L.; Sauerberg, P. 1,2,5-

Thiadiazole Analogues of Aceclidine as Potent m1 Muscarinic Agonists. *J. Med. Chem.* **1998**, *41*, 379−392.
(6) Grigorov, M.; Weber, J.; Tronchet, J. M. J.; Jefford, C. W.; Milhous, W. K.; Maric, D. A QSAR Study of the Antimalarial Activity of Some Synthetic 1,2,4-Trioxanes. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 124−130.
(7) Keller, P. A.; Bowman, M.; Dang, K. H.; Garner, J.; Leach, S. P.; Smith, R.; McCluskey, A. Pharmacophore Development for Corticotropin-Releasing Hormone: New Insights into Inhibitor Activity. *J. Med. Chem.* **1999**, *42*, 2351−2357.
(8) Kaminski, J. J.; Rane, D. F.; Snow, M. E.; Weber, L.; Rothofsky, M. L.; Anderson, S. D.; Lin, S. L. Identification of Novel Farnesyl Protein Transferase Inhibitors Using Three-Dimensional Database Searching Methods. *J. Med. Chem.* **1997**, *40*, 4103−4112.
(9) Sprague, P. W. Automated Chemical Hypothesis Generation and Database Searching with Catalyst. *Perspect. Drug Discovery Des.* **1995**, *3*, 1−20.
(10) Nicklaus, M. C.; Neamati, N.; Hong, H.; Mazumder, A.; Sunder, S.; Chen, J.; Milne, G. W. A.; Pommier, Y. HIV-1 Integrase Pharmocophore: Discovery of Inhibitors through Three-Dimensional Database Searching. *J. Med. Chem.* **1997**, *40*, 920−929.
(11) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.
(12) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117−125.
(13) Barnum, D.; Greene, J.; Smellie, A.; Sprague, P. Identification of Common Functional Configurations Among Molecules. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 563−571.
(14) Guner, O. F. *Pharmacophore Perception, Development, and Use in Drug Design*; IUL: La Jolla, CA, 2000.
(15) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 2. Applications of Conformational Models. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 295−304.
(16) Smellie, A.; Kahn, S. D.; Teig, S. L. Analysis of Conformational Coverage. 1. Validation and Estimation of Coverage. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 285−294.
(17) Smellie, A.; Teig, S. L.; Towbin, P. Poling: promoting conformational variation. *J. Comput. Chem.* **1995**, *16*, 171−187.
(18) Sielecki, T. M.; Boylan, J. F.; Benfield, P. A.; Trainor, G. L. Cyclin-Dependent Kinase Inhibitors: Useful Targets in Cell Cycle Regulation. *J. Med. Chem.* **2000**, *43*, 1−18.
(19) DeBondt, H. L.; Rosenblatt, J.; Jancarik, J.; Jones, H. D.; Morgan, D. O.; Kim, S.-H. Crystal structure of cyclin-dependent kinase 2. *Nature* **1993**, *363*, 595−602.
(20) Jeffrey, P. D.; Russo, A. A.; Polyak, K.; Gibbs, E.; Hurwitz, J.; Massague, J.; Pavletich, N. Mechanism of CDK activation revealed by the structure of a cyclinA-CDK2 complex. *Nature* **1995**, *376*, 313−320.
(21) Schow, S. R.; Mackman, R. L.; Blum, C. L.; Brooks, E.; Horsma, A. G.; Joly, A.; Kerwar, S. S.; Lee, G.; Shiffman, D.; Nelson, M. G.; Wang, X.; Wick, M. M.; Zhang, X.; Lum, R. T. Synthesis and Activity of 2,6,9-Trisubstituted Purines. *Bioorg. Med. Chem. Lett.* **1997**, *7*, 2697−2702.
(22) Furet, P.; Zimmerman, J.; Capraro, H.-G.; Meyer, T.; Imbach, P. Structure based design of potent CDK1 inhibitors derived from olomoucine. *J. Comput.-Aided Mol. Design* **2000**, *14*, 403−409.
(23) Gray, N. S.; Wodicka, L.; Thunniessen, A.-M. W. H.; Norman, T. C.; Kwon, S.; Espinoza, 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.
(24) Carlson, B. A.; Dubay, M. M.; Sausville, E. A.; Brizuela, L.; Worland, P. J. Flavopiridol Induces G1 Arrest with Inhibition of Cyclin-Dependent Kinase CDK2 and CDK4 in Human Breast Carcinoma Cells. *Cancer Res.* **1996**, *56*, 2973−2978.
(25) Vesely, J.; Havlicek, L.; Strnad, M.; Blow, J. J.; Donella-Deana, A.; Pinna, L.; Letham, D. S.; Kato, J. Y.; Detivaud, L.; Leclerc, S.; Meijer, L. Inhibition of Cyclin-Dependent Kinases by Purine Analogues. *Eur. J. Biochem.* **1994**, *224*, 771−786.
(26) Barvian, M.; Boschelli, D. H.; Cossrow, J.; Dobrusin, E.; Fattaey, A.; Fritsch, A.; Fry, D.; Harvey, P.; Keller, P.; Garrett, M.; La, F.; Leopold, W.; McNamara, D.; Quin, M.; Trumpp-Kallmeyer, S.; Toogood, P.; Wu, Z.; Zhang, E. Pyrido[2,3-d]pyrimidin-7-one inhibitors of cyclin-dependent kinases. *J. Med. Chem.* **2000**, *43*, 4606−4616.
(27) Imbach, P.; Capraro, H. G.; Furet, P.; Mett, H.; Meyer, T.; Zimmermann, J. 2,6,9-trisubstituted purines: optimization towards highly potent and selective CDK1 inhibitors. *Bioorg. Med. Chem. Lett.* **1999**, *9*, 91−96.
(28) Kim, K. S.; Sack, J. S.; Tokarski, J. S.; Qian, L.; Chao, S. T.; Leith, L.; Kelly, Y. F.; Misra, R. N.; Hunt, J. T.; Kimball, S. D.; Humphreys, W. G.; Wautlet, B. S.; Mulheron, J. G.; Webster, K. R. Thio- and oxoflavopiridols, cyclin-dependent kinase 1-selective inhibitors: synthesis and biological effects. *J. Med. Chem.* **2000**, *43*, 4126−4134.
(29) Murthi, K. K.; Dubay, M.; McClure, C.; Brizuela, L.; Boisclair, M. D.; Worland, P. J.; Mansuri, M. M.; Pal, K. Structure−activity relationship studies of flavopiridol analogues. *Bioorg. Med. Chem. Lett.* **2000**, *10*, 1037−1041.
(30) Ohlmeyer, M. H. J.; Swanson, R. N.; Dillard, L.; Reader, J. C.; Asouline, G.; Kobayashi, R.; Wigler, M.; Still, W. C. Complex synthetic chemical libraries indexed with molecular tags. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 10922−10926.
(31) De Azevedo, W. F.; Leclerc, S.; Meijer, L.; Havlicek, L.; Strnad, M.; Kim, S. H. Inhibition of cyclin-dependent kinases by purine analogues: crystal structure of human cdk2 complexed with roscovitine. *Eur. J. Biochem.* **1997**, *243*, 518−526.

CI020368A