# Classification of Kinase Inhibitors Using a Bayesian Model

Xiaoyang Xia,*,‡ Edward G. Maliski,‡ Paul Gallant,‡ and David Rogers§

*Amgen, Inc., One Amgen Center Drive, Thousand Oaks, California 91320, and SciTegic, 9665 Chesapeake Drive, Suite 401, San Diego, California 92123*

*Received July 3, 2003*

The use of Bayesian statistics to model both general (multifamily) and specific (single-target) kinase inhibitors is investigated. The approach demonstrates an alternative to current computational methods applied to heterogeneous structure/activity data sets. This approach operates rapidly and is readily modifiable as required. A generalized model generated using inhibitor data from multiple kinase classes shows meaningful enrichment for several specific kinase targets. Such an approach can be used to prioritize compounds for screening or to optimally select compounds from third-party data collections. The observed benefit of the approach is finding compounds that are not structurally related to known actives, or novel targets for which there is not enough information to build a specific kinase model. The general kinase model described was built from a basis of mostly tyrosine kinase inhibitors, with some serine/threonine inhibitors; all the test cases used in prediction were also on tyrosine kinase targets. Confirming the applicability of this technique to other kinase families will be determined once those biological assays become available.

## Introduction

Kinase inhibition is a major focus for therapeutic intervention against a variety of diseases including cancer, inflammatory disorders and diabetes. Several categories of methods have been used to study kinase inhibitors, including 3D methods, clustering and cell-based binning methods, and fitting methods.

As a result of the abundant structural information generated using X-ray and NMR technologies, many computational efforts have involved structure-based design techniques such as ligand−protein docking,[1,2] and three-dimensional (3D) pharmacophoric identification,[3,4] Usually, predictions made with such computational methods were for inhibitors of specific proteins rather than for entire kinase families and were applicable only to proteins with known 3D structure information, which limits their applicability to specific kinases for which the target is known. A number of techniques, such as clustering, binning, and fitting methods, have recently been used to classify compounds without the requirements of 3D protein target information.

Several researchers have used clustering or binning in compound classification. Willett[5] suggested that Jarvis-Patrick clustering was the most effective non-hierarchical method for clustering molecules according to their chemical and biological characteristics. Brown and Martin[6,7] later analyzed the results of different clustering methods and the use of various 2D and 3D descriptors for the classification of active and inactive compounds. They concluded that the combination of a hierarchical Ward's clustering and two-dimensional structural keys performed best. Cell-based partitioning of compounds in lower dimensional chemical space is a method related to clustering, though a grid rather than a distance function is used to bin the samples. Pirard and Pickett[8] used partitioning and BCUTs[9,10] to classify kinase inhibitors active against five different protein kinases.

Fitting methods such as artificial neural networks,[11−13] genetic function approximation (GFA),[14,15] standard and partial least squares,[16−19] and recursive partitioning[20,21] have also been used to analyze structure−activity data or to predict chemical properties. For example, Manallack et al. used a neural network in conjunction with BCUT descriptors to successfully discriminate compounds belonging to a kinase family.[22] The genetic function has been applied widely in compound classification, quantitative structure−activity relationship (QSAR) and quantitative structure−property relationship (QSAP) by a number of investigators.[14,15] Both standard least-squares and partial least-squares analysis (PLS) have a long history of application in quantitative structure−activity relationship (QSAR;,[16−19] in fact, PLS method was incorporated into comparative molecular field analysis (*CoMFA*, a 3D QSAR method).[2325] Stanley Young and colleagues implemented a recursive partitioning method a few years ago that can be used to derive predictive models to distinguish active from inactive compounds.[20,21] This method, compared to some older implementations, has the advantage of being able to handle a large number of descriptors and very large compound data sets.

One is struck by the wide variety of methods available for analysis and the different conclusions reached by different authors as to the preferred method. Because of differences in selected test cases, methods and descriptors, it is difficult to compare directly the performance of the above methods.

Bayesian concepts and methodology has existed for many years; however, its popularity as a tool within

* To whom correspondence should be addressed. Tel: (805) 447-6938. Fax: (805) 499-9955. E-mail: xxia@amgen.com.
‡ Amgen, Inc.
§ SciTegic.

10.1021/jm0303195 CCC: $27.50 
Published on Web 08/04/2004

drug discovery and structure−activity analysis is recent. Therefore, many applications of Bayesian statistics, such as prioritization of compounds for biological testing purchasing, have not been fully explored. Here we report our investigation on a modified Naïve Bayesian statistics, such as the one implemented in SciTegic's Pipeline Pilot, and its application in kinase activity classification. We will describe this process and present several results, building both general (multiclass) and specific (single-target) models of kinase inhibitors.

## Materials and Methods

**Modified Bayesian in Pipeline Pilot.** Bayesian analysis is a statistical categorization method that addresses some of the limitations inherent in conventional fitting methodologies. The learned models are created with a straightforward learn-by-example paradigm: the researcher marks the sample data that is of interest (the "good" samples), and the system learns to distinguish them from other background data. No tuning parameters are required beyond the selection of the input descriptors from which to learn.

The learning process generates a large set of Boolean features from the input descriptors, then collects the frequency of occurrence of each feature in the "good" subset and in all data samples. To apply the model to a particular sample, the features of the sample are generated, and a weight is calculated for each feature using a Laplacian-adjusted probability estimate. The weights are summed to provide a probability estimate, which is a relative predictor of the likelihood of that sample being from the "good" subset.

The Bayesian modeling method offers three important features:

First, it is fast and efficient for large datasets, scaling linearly with respect to the number of molecules. During analysis, data frequency statistics are collected in a single pass and the frequencies themselves become the model. This is in contrast to methods that attempt to fit the data; such methods nearly always scale greater than linearly.

Second, because the method is not a fitting method, it is less affected by the "curse of dimensionality" when large numbers of descriptors are used. This lack of fitting also aids in modeling datasets with extremely small numbers of "good" samples or when there are large sources of noise from false positives and false negatives. The ability to model in a high-dimensional environment also assists modeling structurally dissimilar (noncongeneric) data or in modeling multiple activity classes in a single model (multimodal data).

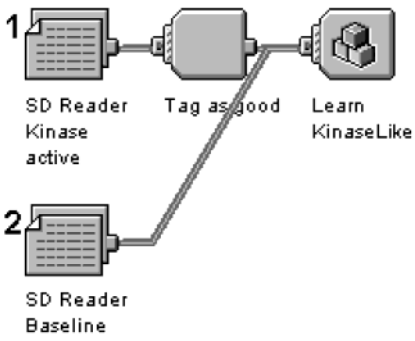
Third, the Bayesian model weights features by assigning greater significance to characteristics that appear to distinguish good samples from baseline samples. This is in contrast to the static distance functions used by clustering methods, such as the Tanimoto Distance between two fingerprints, in which all bits are given equal weight. In such functions, a small number of bits representing features important for activity may be lost among a larger number of bits representing less important features.

**The Laplacian-Corrected Estimator.** The Laplacian-corrected estimator is used to adjust the uncorrected probability estimate of a feature to account for the different sampling frequencies of different features. The derivation is given below.

Assume that $N$ samples are available for training, of which $M$ are "good" (active). An estimate of the *baseline probability* of a randomly chosen sample being active, $P$(Active), is $M/N$.

Next, assume we are given a feature $F$ contained in $B$ samples, and that $A$ of those $B$ samples are active. The uncorrected estimate of activity, $P$(Active|$F$), is $A/B$. Unfortunately, as the number of samples, $B$, becomes small, this estimator tends to be less reliable.

For example, if $A = 1$ and $B = 1$, $P$(Active|$F$) would be 1 (that is, certainly active), which seems overconfident for a feature we have only seen once.




| Parameter Name | Parameter Value |
|---|---|
| LearnedPropertyName | KinaseLike_useall |
| TestForGood | Kinaseactive IS DEFINED |
| UserSoperties | PredefinedSet |
| PredefinedSet | FCFP_6,ALogP,Molecular_Weight,Num_H_Don |
| UserSet | |
| WhenFinished | ShowStatisticsTable |

**Figure 1.** Bayesian categorization protocol in Pipeline Pilot.

Most likely, the estimator is poor because the feature is undersampled, and further sampling of that feature would improve the estimate. We can estimate the effect of further sampling if we assume the vast majority of features have no relationship with activity; that is, if, for most features, $F_i$, we would expect $P$(Active|$F_i$) to be equal to our baseline probability $P$(Active).

If we sampled the feature $K$ additional times, we would expect $P$(Active)*$K$ of those new samples to be active. This provides the information needed to estimate the corrective effect of $K$ additional samples: $P_{corr}$(Active|$F$) = ($A$ + $P$(Active)*$K$)/($B$ + $K$). (For $K = 1/P$(Active), this is the Laplacian correction.)

This correction stabilizes the estimator: as the number of samples, $B$, containing a feature approaches zero, the feature's probability contribution converges to $P$(Active), which would be the expected value for most features.

The final step is to make the estimator a *relative estimate* by dividing through by $P$(Active) − that is, $P_{final}$(Active|$F$) = $P_{corr}$(Active|$F$)/$P$(Active).

For most features, log $P_{final} \sim 0$. For features more common in actives, log $P_{final} > 0$. For features less common in actives, log $P_{final} < 0$. The completed estimate for a particular sample is derived by adding together the log $P_{final}$ values for all the features present in that sample.

**Definition of Datasets.** The following three datasets of Amgen compounds were used in this study:

**CORP**: A dataset of 193 417 Amgen screening compounds. These compounds represent random drug molecules and are referred to in this document as "*baseline*".

**KA**: "Kinase-Active", a subset of the CORP dataset containing 6236 compounds found to inhibit one or more protein kinases with $IC_{50} < 10\ \mu M$. These compounds were tagged as "*good*" molecules.

**SUB**: A subset of the CORP dataset containing 13 253 compounds that are a "preferred list" of compounds used by biologists for kinase assay screening. The bias in this preferred list is based on previous knowledge amassed by chemists and, in cases of vendor-supplied compounds, on information supplied by vendors.

**Building a Bayesian Model.** SciTegic's Pipeline Pilot (version 2.0) was employed to perform the Bayesian analysis to build a model. The computation was undertaken on a four CPU Pentium III server (899 MHz).

Figure 1 shows the building blocks used by Pipeline Pilot to generate the Bayesian Categorization protocol.

A kinase model is generated by first reading in all the "good" (kinase inhibitors) and "baseline" (random drug compounds) molecules, then molecular features and properties are calculated, among them FCFP_6 (functional-class fingerprints, a 2D structural fingerprint where the atom types are abstracted to the role that the atom plays in the molecule, e.g. hydrogen-

bond donor, halogen, or aromatic), AlogP, molecular weight, number of hydrogen-bond donors and acceptors, and number of rotatable bonds.

The software then creates a new "learned property" component that stores the model as a calculator. Unlike regression models, Bayesian does not supply a simple parameter equation. Rather, a cumulative score of feature contributions to "kinase-inhibitor" likeness is computed. Scores must therefore be interpreted by likelihood instead of potency. That is, if compounds A's Bayesian score is 90 and compounds B's score is 70, a correct interpretation is that compounds A is more likely to be an inhibitor than compound B, not that compound A is likely to be more potent than compound B.

This new component can be utilized within another protocol to generate a numerical score that, when applied to a test molecule, can be used to rank the evaluation set on the relative likelihood that the compounds will exhibit the desired activity (in this case, kinase inhibition).

To validate the methodology, the entire CORP dataset was split into a training set and a validation set.

We examined two splits: 1:1 and 1:9. For the 1:1 split, the training set contained 96 628 samples with 3186 known "good" compounds, and the validation set contained 96 789 samples with 3050 known "good" compounds. For the 1:9 split, the training set contained 19 319 samples with 873 known "good" compounds, and the validation set contained 174 098 samples with 5363 known "good" compounds.

The position of active compounds in the ordered list of the validation set is an indicator of the quality of the model. A convenient way to visualize this is by plotting the cumulative distribution of the active compounds in the ranked validation set (called an *enrichment curve*).

Another useful measure of the quality of the model is the enrichment rate in the highest scoring subsets. This is defined as the ratio of the number of actives in a subset to the number of actives in the whole set, divided by the ratio of the number of compounds in the subset to the number of compounds in the whole set.

**Diversity Analysis.** For predictivity over the widest range of structural classes and kinase classes, the predictive model should be built from a heterogeneous dataset, i.e., a dataset of structurally diverse compounds. The KA dataset (containing 6236 "good" compounds) was analyzed to ensure that it consisted of compounds that were dissimilar and that they accurately represented the major structural variations present in the in-house kinase inhibitors and contained examples of inhibitors of all the major kinase classes.

First, in terms of biological diversity, these compounds inhibit as many as 39 protein kinases representing two kinase families (tyrosine and serine/threonine). Second, the chemical diversity was assessed by the following two methods:

**Self-Similarity.**This method employs a distance-based mathematical measurement. Clusters of many highly similar compounds are undesirable. For a given compound *A*, the Tanimoto Distances between *A* and each of the other compounds in the dataset were calculated. The Daylight fingerprint (1024 bits) was used as the structural descriptor. Any compound for which its Tanimoto Distance to *A* was less than a predetermined cutoff value was defined as *A*'s *near neighbor* (NN), and the total number of near neighbors in the vicinity of *A* was calculated. The calculations were performed automatically using a script developed in-house at Amgen.

**Ring Analysis.**This is a visual method for displaying the chemotype within the dataset. The dataset was examined for the number of different ring systems contained within it. Ring fragments were rank-ordered based on their frequency of occurrence using a script originally written by Jerry Young of Daylight[26] and modified by Greg Woo of Amgen. Excel (MS) and JMP4.0 (SAS) were used to analyze the results of the study.

## Results and Discussion

**Diversity Analysis of the Dataset.** Table 1 and Figure 2 show the results of applying Tanimoto Distance cutoffs of 0.15 and 0.25 in Self-Similarity tests performed on the KA set "good" compounds used to build the model (0.15 and 0.25 cutoffs correspond to 85% and 75% similarity, respectively).

**Table 1.** Near Neighbor Statistics of the KA ("Good") Set

| number of near neighbors (NN) | distance cutoff < 0.15 | distance cutoff < 0.25 |
|---|---|---|
| 0 | 1824 | 1361 |
| 1−2 | 1451 | 872 |
| 3−4 | 665 | 717 |
| 5−9 | 765 | 715 |
| ≥10 | 1531 | 2571 |

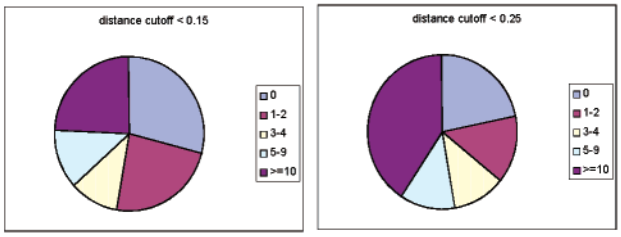



**Figure 2.** Number of near neighbors distribution of the KA ("good") set.

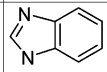
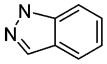
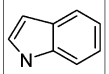
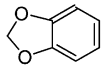
The results show that (1) using either cutoff, less than 50% of the compounds have five or more neighbors, and (2) depending on the cutoff, 30−40% of the compounds are highly similar (NN ≥ 10). This similarity was attributed to the fact that Amgen's in-house medicinal chemistry project teams synthesized many of the compounds.

In addition to examining the numbers for Tanimoto Distance calculation, the KA dataset was broken down to fragment level to ensure structural dissimilarity of the compounds. Since rings are often good representations of chemical families, a ring analysis script was used to extract rings from each molecule. The script searched for larger rings (e.g. indoles) and ignored smaller component rings (e.g. benzene and pyrrole). The script extracted 573 rings; the 20 rings occurring with the highest frequency are listed in Table 2. The table also shows smaller and less interesting rings (such as pyrimidine/furan) as well as larger and more interesting rings (such as indole/indazole).

The results from Tanimoto Distance calculations and from ring analyses show that the KA ("good") set contains diverse molecules including a variety of structure motifs, which suggests that the compounds are good candidates for the kinase-learning test.

**Validation of the Kinase Model.** Figure 3 is an *enrichment curve* generated by Pipeline Pilot and is a visual display of the enrichment obtained using the 1:1 split to predetermine the "good" compounds in the validation set. In a random, unbiased screening, plotting the percentage of "good" compounds found (*Y*-axis) against the percentage of compounds screened (*X*-axis) should yield a constant and that is clearly shown by the straight line. If the samples are rank-ordered according to the their likelihood of exhibiting activity and then screened, the active compounds should be found more rapidly than if they are screened at random and the plot appears as the curve shown. Figure 4 shows the enrichment rate (*Y*-axis) against the percentage of compounds screened (*X*-axis), the reciprocal of the enrichment plot shown in Figure 3.

**Table 2.** Twenty Most Frequently Occurring Rings in the KA ("Good") Set

| Frequency of Occurrence | Ring | Frequency of Occurrence | Ring | Frequency of Occurrence | Ring | Frequency of Occurrence | Ring |
|---|---|---|---|---|---|---|---|
| 1081 |  | 105 |  | 34 |  | 27 |  |
| 370 |  | 67 |  | 32 |  | 22 |  |
| 139 |  | 60 |  | 29 |  | 20 |  |
| 119 |  | 52 |  | 27 |  | 20 |  |
| 109 |  | 51 |  | 27 |  | 19 |  |

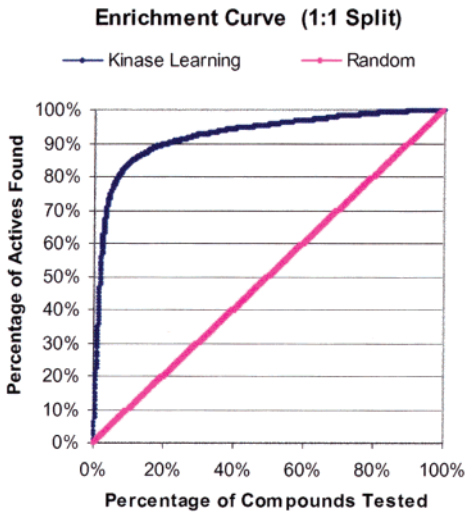



**Figure 3.** Enrichment curve. Kinase-1:1 split model versus random screening of the validation set.

Figure 3 shows that 85% of the "good" compounds occurred in the top 10% of the ordered compounds, which corresponds to an 8-fold enrichment compared with random screening. This illustrates the power of the Bayesian-derived model to capture characteristics about general kinase inhibitors that aid in their detection in the validation set.

Subsequently the CORP dataset was randomly split 1:9 to examine whether a smaller training set would have similar predictive ability. A second model was built based on a training set of 10% of CORP and then used to predict the results for the remaining 90%.

The results from Validation Set 2 are shown in Figure 5 and Figure 6 together with the results from the Validation Set from the 1:1 split described earlier. The results suggest that the kinase model built using only 10% of the data is nearly as good as the model built with the larger training set, and that the use of 50% of the data for learning is probably redundant.

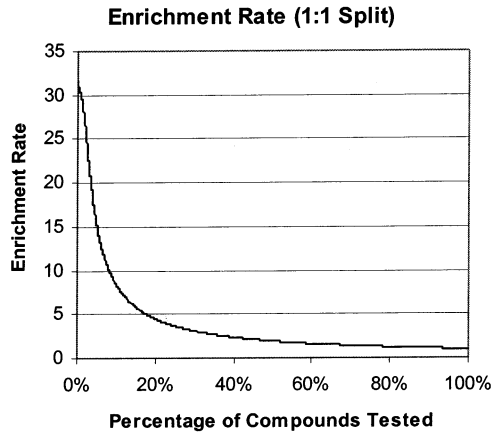



**Figure 4.** Enrichment rate using the kinase-learning model of the validation set.

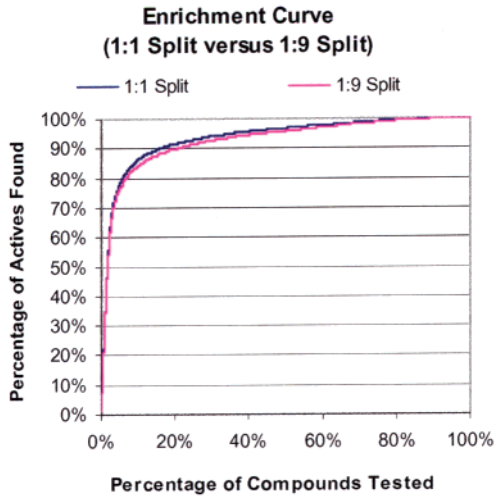



**Figure 5.** Enrichment Curve. 1:1 Split model versus 1:9 split model.

From Table 1, since only 20−30% of the KA set "good" compounds are unique and 70−80% have at least one neighbor, a question arises: could the easy identification

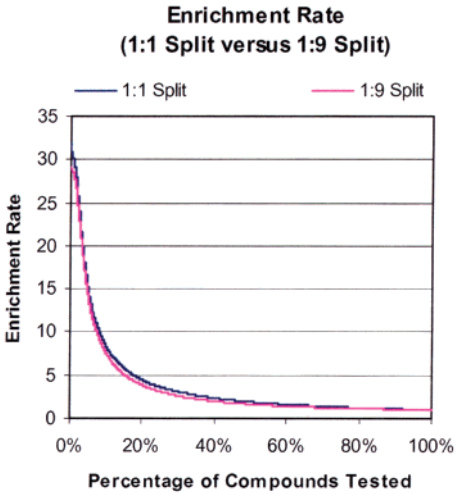



**Figure 6.** Comparison of enrichment rate. 1:1 Split model versus 1:9 split model.

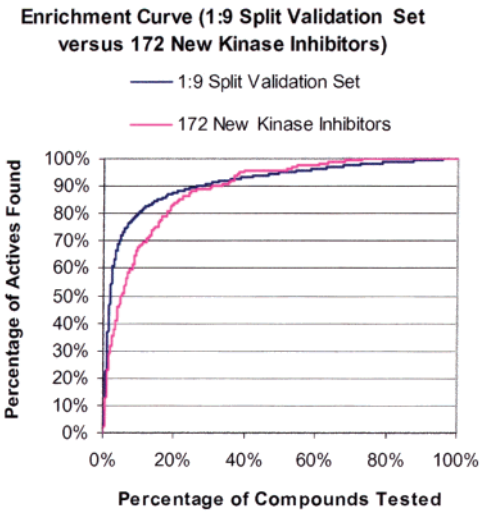



**Figure 7.** Enrichment curve using 1:9 split model to identify new classes of kinase inhibitors and in-house kinase inhibitors.

of "good" compounds be simply due to the similarity of the compounds? Stated differently, might the Bayesian analysis also be effective in identifying new classes of kinase inhibitors that are markedly different from those used in the learning set, considering the fact that many modeling techniques only select compounds that resemble the training set, thus making it potentially difficult to avoid patent problems if these compounds are used as starting points for optimization?

To attempt to answer this question, 172 newly identified compounds, described in recent literature as newly emerging protein kinase inhibitor classes,[3] were chosen for additional testing. These compounds were specifically selected to *not* overlap with the classes represented by the KA set; all had Tanimoto Distances > 0.40 from all the KA set compounds.

The set of 172 newly identified compounds was merged with the 1:9 validation set, then all KA compounds were removed, which yielded 168 907 samples with 172 "good" compounds. If we apply the 1:9 split model to this set, can we identify the 172 kinase inhibitors quickly, and if so, how does the performance compare with the 1:9 validation set, which only contains in-house known "good" compounds?

Figure 7 depicts the enrichment curve. Although these compounds represent new kinase inhibitor classes, we are still able to find 70% of them in the top 10% rank-ordered compounds and 85% in the top 20% rank-ordered compounds. The active compounds discovery rate is very similar and only slightly worse than finding the in-house kinase inhibitors.

This demonstrates that in addition to identifying "good" compounds, the Bayesian analysis was also able to identify compounds from novel kinase and structural classes. This offers a significant advantage over many other classification methods.

**Applications.** To test the modeling method in the real world, we applied it to two specific applications in order to answer the following questions:

1. Can the model help us rank order the screening of compounds so that the efficiency of a screening is improved?

2. Can the model be used in the design of a kinase-preferred collection that can be used to frontload any kinase assay?

**Rank Order Screening of Compounds.** The results of the validation tests described earlier show that the kinase model is efficient in identifying compounds that have a high likelihood of producing a hit in a kinase assay. A complete kinase model was thus generated using all data (KA as "good" and CORP as "baseline"); all CORP compounds were assigned a score and rank-ordered. The score is an indication of the likelihood of the given compound to inhibit a kinase, but is not an indicator of the potency.

Approximately 90% of the "good" compounds in the KA set occurred within the top 10% of the rank-ordered CORP set list (corresponding to approximately 20 000 compounds). We feel that ranking in the top 20 000 compounds is probably an appropriate cutoff value to decide whether a compound is worth screening.

Our biologist colleagues agreed to use the score computed from the model to prioritize a panel of 22 new assays (against which KA set compounds had not yet been tested). Since they would be running a considerable number of assays, screening all 193 417 CORP compounds was considered an onerous task. Instead it was decided to screen SUB, a biased subset of 13 253 CORP compounds which consisted largely of compounds synthesized in-house by kinase project teams, and vendor compounds purchased with kinase in mind.

The histogram and quantiles table in Figure 8 show the distribution of the rank orders of the SUB set compounds among all CORP compounds using the kinase model. According to the quantiles table, nearly 50% of the SUB set compounds fell in the top 20 000 ranked compounds.

The results of applying 22 kinase assays against compounds in the SUB set are shown in Table 3 (the results define a hit as POC < 50). For each assay, the following data are shown:

1. Total number of hits ($a$) occurring with compounds in the entire SUB set.

2. Number of hits ($b$) that occurred using compounds with rank order <20 000 in SUB set.

3. Percentage ratio of hits obtained: $b/a$.

4. Percentage ratio of hits missed: $1 - (b/a)$.

Table 3 shows that, for most assays, screening compounds that are rank-ordered between 1 and 20 000

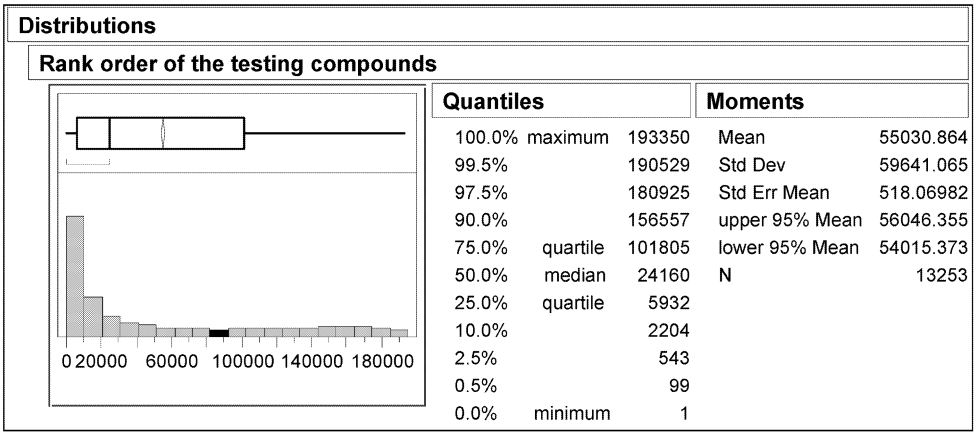



**Figure 8.** Rank order histogram of the SUB Set of CORP using the kinase model.

**Table 3.** Screening Result of the SUB Set in 22 Kinase Assays

| assay | total number of hits POC < 50 in SUB | number of hits POC < 50 and rank < 20 000 | percentage of hits with rank < 20 000 | false negative rate |
|---|---|---|---|---|
| 1 | 286 | 230 | 80.42 | 19.58 |
| 2 | 918 | 823 | 89.65 | 10.35 |
| 3 | 254 | 235 | 92.52 | 7.48 |
| 4 | 282 | 270 | 95.74 | 4.26 |
| 5 | 797 | 755 | 94.73 | 5.27 |
| 6 | 169 | 168 | 99.41 | 0.59 |
| 7 | 652 | 642 | 98.47 | 1.53 |
| 8 | 44 | 29 | 65.91 | 34.09 |
| 9 | 582 | 503 | 86.43 | 13.57 |
| 10 | 806 | 776 | 96.28 | 3.72 |
| 11 | 166 | 146 | 87.95 | 12.05 |
| 12 | 625 | 589 | 94.24 | 5.76 |
| 13 | 387 | 381 | 98.45 | 1.55 |
| 14 | 280 | 267 | 95.36 | 4.64 |
| 15 | 424 | 375 | 88.44 | 11.56 |
| 16 | 591 | 535 | 90.52 | 9.48 |
| 17 | 592 | 561 | 94.76 | 5.24 |
| 18 | 94 | 91 | 96.81 | 3.19 |
| 19 | 194 | 187 | 96.39 | 3.61 |
| 20 | 77 | 73 | 94.81 | 5.19 |
| 21 | 682 | 679 | 99.56 | 0.44 |
| 22 | 663 | 641 | 96.68 | 3.32 |

(one-half the SUB set) yields over 90% of the total hits (using POC < 50 to define a hit). The percentage varied from 66% to 99% with a mean of 93%. This indicates that on average 93% of hits are identified (only 7% are false negatives) by a limited screening of the 50% of the SUB compounds that are ranked below 20 000. This demonstrates that the general kinase model did act as an effective method of prioritizing screening compounds within the kinases tested.

**Kinase-Preferred Collection**. Biologists often start with a *preferred set* comprising a few hundred to a few thousand compounds in order to validate an assay, or more frequently, as the initial set of a sequential screening process.

Recently the work on sequential screening (also called *smart screening*) has been reviewed by both Young[27] and Engels.[28] In this approach, a relatively small number of compounds (the initial set) is screened and the results are analyzed statistically to produce a mathematical model. The model is used to select additional compounds for screening. Such a strategy has often been effective in exploiting the potential of HTS in smarter and more cost-efficient ways.

**Table 4.** Bilevel Clustering Method for Selecting a Kinase-Preferred Set

| compound ID | fingerprint-based clustering | BCUT-based binning |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 1 | 2 |
| 3 | 1 | 2 |
| 4 | 2 | 1 |
| 5 | 2 | 2 |
| 6 | 3 | 4 |
| 7 | 3 | 4 |
| 8 | 3 | 5 |

Ideally, compounds in the initial set should be defined by the following criteria:

1. Compounds must have a higher likelihood of binding to the target.

2. Compounds must be structurally diverse.

3. Compounds should not be isolated but somehow related to other members of the preferred set, so that after performing biological experiments, structure−activity knowledge can be extracted from the results.

The derived kinase model was used to help construct the kinase-preferred set. As discussed earlier, limiting compounds to those occurring within the top 20 000 of the rank-ordered set satisfies criterion no. 1. However, screening against 20 000 compounds still requires considerable effort and may not be feasible for low throughput screening tests (i.e. dose−response). Additional sampling and paring down of the compound test set are required to satisfy criterion no. 2 and 3.

A bilevel clustering method was used for this purpose. Fingerprint-based clustering was utilized to sample the chemical families, and within each chemical family the physical property space was sampled using BCUT-based binning technology.[9] Using this method, selected compounds are diverse yet are related to each other when sharing the same cluster/bin membership (shaded rows in Table 4).

Finally, an analytical experiment was conducted to ensure the selected compounds were of high quality. This selection method yielded 972 compounds to form a kinase-preferred set.

Once the kinase-preferred set is in place, it is used as the initial set of a sequential screening process. Figure 9 illustrates our sequential screening paradigm. The screening results from the kinase-preferred set are analyzed to produce a second and more refined Bayesian model with information specific only to the particular

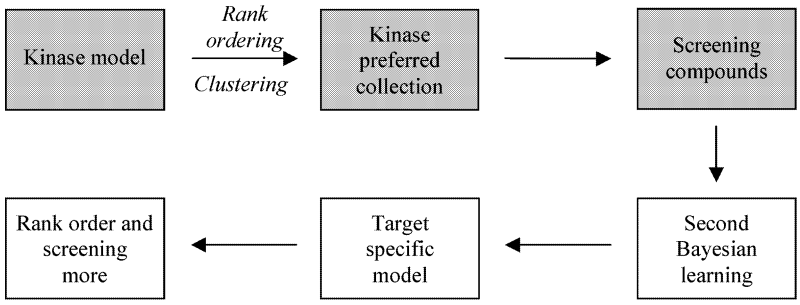



**Figure 9.** Sequential screening compound flow.

**Table 5.** Screening Results of the Kinase-Preferred Set in 16 Kinase Assays

| kinase binding assay ID | number of hits with $IC_{50} < 10\ \mu M$ | hit rate |
|---|---|---|
| 1 | 39 | 4 |
| 2 | 82 | 8 |
| 3 | 52 | 5 |
| 4 | 107 | 11 |
| 5 | 33 | 3 |
| 6 | 29 | 3 |
| 7 | 66 | 7 |
| 8 | 12 | 1 |
| 9 | 123 | 13 |
| 10 | 29 | 3 |
| 11 | 45 | 5 |
| 12 | 61 | 6 |
| 13 | 40 | 4 |
| 14 | 130 | 13 |
| 15 | 116 | 12 |
| 16 | 127 | 13 |
| average | 68 | 7 |

protein inhibitors instead of to the entire kinase family inhibitors. This information is then used to predict the compounds that are more likely to bind to that protein and that are to be screened in the second iteration.

Following the paradigm, 16 tyrosine kinase assays were front-loaded with the kinase-preferred set and their activities were carefully measured in terms of $IC_{50}$. The results are illustrated in Table 5.

The table shows that the hit rates from the kinase-preferred set ranged from 3−13% with an average of 7%. Although it is a marked improvement on the average hit rate of 0.1% observed with random screenings at Amgen, the hit rate of the second screening iteration jumps significantly as we use the information obtained from the initial screening to extract further knowledge *specific to the target* to prioritize the screening compounds.

For example, in assay 1, the 39 compounds identified with $IC_{50} < 10\ \mu M$ were used to build a "target 1-specific" learning model. The model was then applied to the entire CORP set and suggested the next 100 compounds which have the highest likelihood of binding to target 1. When biologists tested the top 100 suggested compounds, 43 were observed with $IC_{50} < 10\ \mu M$.

The same experiment was repeated on assay 3. Knowledge gained from the initial screening recommended 100 top ranked compounds for screening, and 47 were found with $IC_{50} < 10\ \mu M$. Unfortunately, due to the limited biology resources, this experiment was only performed on the two assays referred to here and not across all 16 assays, but we have been able to prove the concept.

## Conclusion

In this study we investigated the use of Bayesian statistics to model both general (multifamily) and specific (single-target) kinase inhibitors. This approach appears to give useful insight for the kinase activity classification problem. The method did operate rapidly and is readily modifiable as required.

Building the model based on a 200 000 compound set took three minutes, and evaluation of the 10 000 compound library took 10 s. For two examples, the general kinase model demonstrated meaningful enrichment for several specific kinase targets, relative to random screening, without having to develop specific models for inhibitors of each individual kinase. Such an approach can be used to prioritize the compounds for screening or to optimally select compounds from third-party data collections. One observed benefit of the approach is the possibility of finding compounds that are not structurally related to known actives, or novel targets for which there is not enough information to build a specific kinase model.

The general kinase model described in this paper was built from a basis of mostly tyrosine kinase inhibitors (~80%); all the test cases were also on tyrosine kinase inhibitors. Confirming the applicability of this technique to other kinase families will be determined once those biological assays become available.

**Acknowledgment.** This work was supported by research funding generously provided by Amgen Inc. (Thousand Oaks). The authors thank Greg Woo for writing/modifying the scripts for the near neighbor comparison and ring analysis. We also thank Mr. Peter Brooks and Mr. Ken Kashani for their help in preparing the manuscript.

## References

(1) Woolfrey, J. R.; Weston, G. S. The Use of Computational Methods in the Discovery and Design of Kinase Inhibitors. *Curr. Pharm. Des.* **2002**, *8*, 1527−1545.
(2) Majeux, N.; Scarsi, M.; Tenette-Souaille, C.; Caflisch, A. Hydrophobicity Maps and Docking of Molecular Fragments With Solvation. *Perspect. Drug Discovery Des.* **2000**, *20*, 145−169.
(3) Dumas, J. Protein Kinase Inhibitors: Emerging Pharmacophores 1997−2000. *Exp. Opin. Ther. Pat.* **2001**, *11*, 405−429.
(4) Toledo, L. M.; Lydon, N. B.; Elbaum, D. The Structure-Based Design of ATP−Site Directed Protein Kinase Inhibitors. *Curr. Med. Chem.* **1999**, *6*, 775−805.
(5) Willett, P.; Wintermann, V.; Bawden, D. Implementation of Nonhierarchic Cluster Analysis Methods in Chemical Information System: Selection of Compounds for Biological Testing and Clustering of Substructure Search Output. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 109−118.
(6) Brown, R. D.; Martin, Y. C. Use of Structure−Activity Data to Compare Structure Based Clustering Methods and Descriptors for Use in Compound Selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.

(7) Brown, R. D.; Martin, Y. C. The Information Content of 2D and 3D Structural Descriptors Relevant to Ligand−Receptor Binding. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 1−9.
(8) Pirard, B.; Picket, S. D. Classification of Kinase Inhibitors Using BCUT Descriptors. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1431−1440.
(9) Pearlman, R. S.; Smith, K. M. Novel Software Tools for Chemical Diversity. *Perspect. Drug Discovery Des.* **1998**, *9*, 339−353.
(10) Stanton, D. T. Evaluation and Use of BCUT Descriptors in QSAR and QSAP Analysis. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 11−20.
(11) Gasteiger, J.; Zupan, J. Neural Networks in Chemistry. *Angew. Chem., Int. Ed. Engl.* **1993**, *32*, 503−536.
(12) Burden, F. R. Using Artificial Neural Networks to Predict Biological Activity from Simple Molecular Structural Considerations. *Quant. Struct.-Act. Relat.* **1996**, *15*, 7−11.
(13) King, R. D.; Hirst, J. D.; Sternberg, M. J. E. New Approaches to QSAR: Neural Networks and Machine Learning. *Perspect. Drug Discovery Des.* **1993**, *1*, 279−290.
(14) Rogers, D.; Hopfinger, A. J. Application of Genetic Function Approximation to Quantitative Structure−Activity Relationships and Quantitative Structure−Property Relationships. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 854−866.
(15) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(16) Martin, Y. C. *Quantitative Drug Design*; Marcel Dekker: New York. 1978.
(17) Ramsden, C. A. Quantitative Drug Design. In *Comprehensive Medicinal Chemistry*; Hansch, C., Ed.; Pergamon Press: New York, 1990; Vol. 4.
(18) Dunn, W. J. Quantitative Structure−Activity Relationships. In *Drug Discovery Technologies*; Clark, C. R.; Moos, W. H., Ed.; Ellis Horwood Limited: New York, 1990; Chapter 2.
(19) Zhu, L. L.; Hou, T. J.; Chen, L. R.; Xu, X. J. 3D QSAR Analysis of Novel Tyrosine Kinase Inhibitors Based on Pharmacophore Alignment. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1032−1040.
(20) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(21) Chen, X.; Rusinko, A., III; Young, S. S. Recursive Partitioning Analysis of a Large Structure−Activity Data Set Using Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.
(22) Manallack, D. T.; Pitt, W. R.; Gancia, E.; Montana, J. G.; Livingstone, D. J.; Ford, M. G.; Whitley, D. C. Selecting Screening Candidates for Kinase and G Protein-Coupled Receptor Target Using Neural Networks. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 1256−1262.
(23) Cramer, R. D. Patterson, D. E.; Bunce, J. D. Comparative Molecular Field Analysis (*CoMFA*). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.
(24) Clark, M.; Cramer, R. D.; Jones, D. M.; Patterson, D. E.; Simeroth, P. E. Comparative Molecular Field Analysis (*CoMFA*). 2. Toward its Use With 3-D Structural Databases. *Tetrahedron Comput. Methodol.* **1990**, *3*, 47−59.
(25) Cramer, R. D.; Wold, S. B. US Patent 5,025, 388, June 18, 1991.
(26) http://www.daylight.com/support.
(27) Young, S. S.; Lam, R. L.; Welch, W. J. Initial Compound Selection for Sequential Screening. *Curr. Opin. Drug Discovery Dev.* **2002**, *5*, 422−427.
(28) Engels, M. F. M.; Venkatarangan, P. Smart Screening: Approach to Efficient HTS. *Curr. Opin. Drug Discovery Dev.* **2001**, *4*, 275−283.

JM0303195