Proteochemometric Recognition of Stable Kinase Inhibition Complexes Using Topological Autocorrelation and Support Vector Machines

Michael Fernandez,[†] Shandar Ahmad,[‡] and Akinori Sarai*,[†]

Department of Bioscience and Bioinformatics, Kyushu Institute of Technology (KIT), 680-4 Kawazu, Iizuka, 820-8502 Japan, and National Institute of Biomedical Innovation, 7-6-8, Saito-Asagi, Ibaraki-shi, Osaka 5670085, Japan

Received February 6, 2010

Intensive research has been performed on computational design of kinase inhibitors using molecular dynamics simulations, docking and quantitative structure−activity relationship (QSAR) analyses, all of which have their own limitations. In this paper, we report the application of proteochemometrics, a ligand-target modeling approach, to the recognition of stable and unstable kinase-inhibitor complexes using support vector machines (SVM) classifiers. The algorithm consists of creating topological autocorrelation descriptors for kinases and inhibitors and then development of SVM models to relate the feature vectors to the stability class (stable or unstable) of hypothetical protein-inhibitor complexes. The approach based on the autocorrelation features was compared with fragment-based approach and the former was found to outperform the later. The final classifier could recognize 82% of data to be stable or unstable using jackknife type of validation and test set prediction. Analysis of substructure classification showed a very homogeneous behavior of the model on the whole target-ligand space. The predictor is available online at http://gibk21.bse.kyutech.ac.jp/AUTOkinI/SVMpredictor.html.

1. INTRODUCTION

Death by cancer has recently increased all around the world.[1] For example, in US this illness has surpassed heart disease as the leading cause of death for people under 85 years age.[1] It has been reported that the improvements in conventional cancer treatments, such as surgery, radiation, and cytotoxic chemotherapy will not substantially impact the clinical outcomes for cancer patients in the future.[1] Because of this, research has focused on alternative clinical strategy, such as the development of a variety of protein-targeted molecule-based cancer therapies, especially selective kinase inhibitors.[1] Need for the discovery of novel kinase inhibitors has therefore acquired a particular significance and, in turn, has led to more basic efforts at the understanding of kinase inhibition process and methods to predict the stability of a potential kinase−inhibitor complex. There are approximately 500 kinases encoded in the human genome having potential role in cancer.[2−4] A better understanding of kinase−substrate and kinase−inhibitor interactions at the molecular level is emerging and continues to be explored. The number of available high-resolution X-ray crystal structures of kinase-inhibitor complexes has substantially increased during recent years. Structural information obtained from these complexes has become an important guide in designing selective potential kinase inhibitors. It is of utmost importance to make the best use of available structure information from these complexes in order to predict the behavior of hypothetical complexes as an aid for inhibitor design.[5,6]

The use of computer-aided design methods can help in extracting the structural features and binding characteristics of the active sites of kinases and ultimately help in minimizing kinase inhibitor-related side effects of drugs by increasing specificity.[7−9] Molecular dynamics and docking-type techniques have helped in elucidating the structural diversity of kinases and their specific interactions with inhibitors.[10−19] On the other hand, quantitative structure−activity relationship (QSAR) studies have been successfully applied to the modeling of inhibitor activities.[20−32] In addition, protein QSAR models have been developed to predict enzyme activity from protein 3D structure.[33−37] However, additional efforts are needed to take the results of computational studies to the level of experimental accuracy, both to provide a screened set of compounds and predict the outcome of experiments.

In this paper, we have applied structural fragments and topological autocorrelations approaches to a ligand data set with known inhibitory activities toward 62 kinases. A target-ligand approach named proteochemometrics (PCM)[38] was employed for the development of classification models of kinase inhibition. The structural information of the target kinase was encoded in sequence fragments (SF) descriptors and amino acids sequence autocorrelation (AASA) vectors.[39−42] Similarly, the structure information of ligands was extracted by fragment and topological approaches. Subsequently, support vector machines (SVMs) model is used to classify hypothetical complexes into stable and unstable classes.

2. MATERIALS AND METHODS

2.1. Data Set. The kinase inhibition data: A total of 8235 inhibitors for 95 sequences of kinase, was obtained from our

* To whom correspondence should be addressed. Tel: 81-948-29-7811. Fax: 81-948-29-7841. E-mail: sarai@bio.kyutech.ac.jp.
[†] Kyushu Institute of Technology.
[‡] National Institute of Biomedical Innovation.

10.1021/ci1000532
Published on Web 06/04/2010

in-house manually curated and annotated protein−ligand interaction data in ProLINT database.[43] Annotations include comprehensive information about experimentally determined thermodynamic, structural, clinical and activity parameters. Kinase sequences were retrieved from UniProt database[44] and added to the kinase inhibition data set. Instant JChem software[45] was used for chemical database management. In many cases, ProLINT data do not contain or have ambiguous values for some of the parameters for a given interaction, as it may have not been reported in the corresponding literature source. The data set was therefore filtered according to the following criteria: (1) inhibitors reporting IC_{50} (81% of kinase entries in ProLINT contain IC_{50}, 5.7% for K_i, and 13% for percent of inhibitor activity), (2) inhibitors fulfilling the mass bioavailability constraint (molecular weight <500 g/mol), and (3) inhibitors reporting unambiguous sequence information for the kinase on which experiment was performed.

After the filtering process, the redundant entries were removed and finally a data set of 3595 nonredundant inhibition complexes of 2233 unique inhibitors with 62 kinases from 19 kinase families was selected (data available from the authors upon request). Inhibition complexes were labeled into two classes according to the affinity threshold of 1 μM: "stable" class ($IC_{50} < 1$ μM) and "unstable" class ($IC_{50} > 1$ μM) yielding 1200 stable and 2395 unstable complexes. Structural fragment (SF) descriptors and 2D autocorrelation vectors for the ligands were computed by Dragon software.[46] Sequence structural fragment (SSF) descriptors were implemented in Matlab[47] environment, and AASA vectors were computed by Protmetrics in-house software.[48] The data set was divided into training (80% data set) and test sets (20% data set) by k-means clustering. Five clusters were generated and cases were homogeneously added to training and test sets by selecting instances from each cluster according to cluster's sizes. The data set is available as Supporting Information. Predictor's optimization was carried out by 3-fold-out (TFO) crossvalidation. The training set was divided into three subsets: two subsets were used for training the classifier and the rest subset was then predicted. This process was repeated until all the subsets were predicted.

2.2. Proteochemometrics (PCM) Modeling. PCM proposed by Wikberg,[38] originates from chemometrics, the mathematical methods to analyze chemical data. PCM models describe the interactions between a set of macromolecules (such as proteins) and a series of ligands. These models are useful for predicting the affinities of new proteins for their ligands if the new molecules fall within the descriptor space of the protein−ligand pairs of the training data set. Similarly, one PCM model can predict the affinity of new ligands toward a group of related targets. A PCM experiment is typically described by three descriptor blocks; the ligand descriptor, protein descriptor, and ligand-protein cross-term blocks. A vector of variables of ligand descriptors characterizes each ligand. Similarly, each protein is described by protein descriptors. Depending on the problem, one or more descriptor blocks can be discarded. In our study, the cross-term blocks were discarded since nonlinearity was automatically incorporated into the models by a nonlinear kernel function. SVMs were trained with a feature matrix obtained by simple concatenation of the target descriptor and protein descriptor blocks.

2.2.1. Structural Fragments Approach. 2.2.1.1. Structural Fragments (SF) Descriptors. SF were computed for ligands by counting 120 fragments in the chemical structures by Dragon computer software.[46]

2.2.1.2. Sequence Structural Fragments (SSF) Descriptors. SF descriptors were computed for the 20 amino acids by Dragon computer software[46] and relative amino-acid compositions were computed for the kinase sequences by a Matlab[47] code. Afterward, sequence structural fragments (SSF) descriptors (120 × 1 row vector) were calculated in Matlab[47] for each kinases as the matrix product of the amino-acid SF descriptors (120 × 20 matrix) by the relative amino-acid composition of kinases (20 × 1 column vector).

2.2.2. Topological Autocorrelation Approach. 2.2.2.1. 2D Spatial Autocorrelation Vectors. The binding of a ligand to a target depends on the shape of the ligand and on a variety of factors such as molecular electrostatic potential, polarizability, hydrophobicity, and lipophobicity. Therefore, in a QSAR study the strategy for encoding molecular information, in some way, either explicitly or implicitly, should account for these physicochemical effects. Furthermore, data sets usually include molecules of different size with different numbers of atoms, so the structural encoding schemes must allow comparing such molecules. Thus, we were faced with the problem of having to compare molecules with different numbers of atoms. Information with variable length can be transformed into fixed-length information by autocorrelation.[49]

Autocorrelation vectors have several useful properties. First, a substantial reduction in data can be achieved by limiting the topological distance, l. Second, the autocorrelation coefficients are independent of the original atom numbering, so they are canonical. And third, the length of the correlation vector is independent of the size of the molecule.[49]

For the autocorrelation vectors, H-depleted molecular structure is represented as a graph and physicochemical properties of atoms (i.e., atomic van der Waals volumes, atomic Sanderson electronegativities, and atomic polarizabilities) as real values assigned to the graph vertices.

These descriptors can be obtained by summing up the products of certain properties of two atoms, located at given topological distances or spatial lag in the molecular graph. Broto-Moreau's autocorrelation vectors were employed for encoding the topological structure of the kinase inhibitors.

Broto-Moreau's autocorrelation coefficient[49] is defined as

$$\mathrm{ATS}lp_k = \sum_{ij} \delta_{ij} p_{ki} p_{kj} \tag{1}$$

where $\mathrm{ATS}lp_k$ is Broto-Moreau's autocorrelation coefficient at spatial lag l, p_{ki} and p_{kj} are the values of property k of atom i and j, respectively, and $\delta(l,d_{ij})$ is a delta function defined as

$$\delta(l, d_{ij}) = \begin{cases} 1 & \text{if } d_{ij} = l \\ 0 & \text{if } d_{ij} \neq l \end{cases} \tag{2}$$

where d_{ij} is the topological distance or spatial lag between atoms i and j.

Dragon computer software[41] was used for calculating the 2D autocorrelation vectors at spatial lags ranging from 1 to 8 and weighted by 3 atomic properties: atomic van der Waals

volumes, atomic Sanderson electronegativities and atomic polarizabilities, thus a total of 24 (8 × 3) 2D autocorrelation vectors were computed.

2.2.2.2. Amino-Acid Sequence Autocorrelation (AASA) Vectors. Autocorrelation vector formalism can be easily extended to amino-acid sequences considering protein primary structure as a linear graph with nodes formed by amino-acid residues. We recently introduced the AASA vectors for modeling the functional variations upon mutation of the ghrelin receptor[39] and the conformational stability of human lysozyme,[40] gene V protein,[41] and chymotrypsin inhibitor 2 mutants.[42] The calculated autocorrelation vectors encode information concerning whole protein sequence. Particularly, AASA vectors of lag l are calculated as follows:

$$\mathrm{AASA}lp_k = \frac{1}{L}\sum_{ij}\delta_{ij}p_{ki}p_{kj} \quad (3)$$

where $\mathrm{AASA}lp_k$ is the AASA at spatial lag l weighted by the p_k property, L is the number of elements in the sum, p_{ki} and p_{kj} are the values of property k of amino acids i and j in the sequence, respectively, and $\delta(l, d_{ij})$ is the delta function in eq 2.

For example, if we consider the decapeptide ASTCGFHCSD, AASA vectors at spatial lag 1 and 5 are calculated as follows:

$$\mathrm{AASA}1p_k = \frac{1}{9}(p_{kA}\cdot p_{kS} + p_{kS}\cdot p_{kT} + p_{kT}\cdot p_{kC} + p_{kC}\cdot p_{kG} + p_{kG}\cdot p_{kF} + p_{kF}\cdot p_{kH} + p_{kH}\cdot p_{kC} + p_{kC}\cdot p_{kS} + p_{kS}\cdot p_{kD}) \quad (4)$$

$$\mathrm{AASA}5p_k = \frac{1}{5}(p_{kA}\cdot p_{kF} + p_{kS}\cdot p_{kH} + p_{kT}\cdot p_{kC} + p_{kC}\cdot p_{kS} + p_{kG}\cdot p_{kD}) \quad (5)$$

In a protein, autocorrelation analysis tests whether the value of a property at one residue is independent of the values of the property at neighboring residues. If dependence exists, the property is said to exhibit spatial autocorrelation. *AASA* vectors represent the degree of similarity between amino acid sequences.

As weights for sequence residues were employed, 7 physicochemical and conformational amino acid/residues properties (Table 1SI in Supporting Information) selected from the AAindex database.[50] In our work, the spatial lag, l, ranged from 1 to 5. AASA vectors were calculated by Protmetrics in-house software.[48] A data matrix of 35 AASA vectors, 7 properties ×5 different lags, was generated with the autocorrelation vectors calculated for each kinase.

2.4. Support Vector Machine (SVM). SVM is a machine learning method, which has been used for many kinds of pattern recognition problems.[51] Since there are excellent introductions to SVMs[51] only the main idea of SVMs applied to pattern classification problems is described here. First, the input vectors are mapped into one feature space (possible with higher dimensions). Second, a hyperplane which can separate two classes is constructed within this feature space. Only relatively low-dimensional vectors in the input space and matrix products in the feature space will be involved in the mapping function. SVM was designed to minimize structural risk whereas previous techniques were usually based on minimization of empirical risk. SVM is usually less vulnerable to the overfitting problem and it can deal with a large number of features.

The mapping into the feature space is performed by a kernel function. There are several parameters in the SVM, including the kernel function and regularization parameter. The kernel function and its specific parameters, together with regularization parameter, cannot be set from the optimization problem but have to be tuned by the user. These can be optimized by the use of Vapnik-Chervonenkis bounds, crossvalidation, an independent optimization set, or Bayesian learning. In this paper, the Radial Basic Function (RBF) was used as kernel function. A grid search was implemented for setting two SVM's parameters, regularization parameter (C) and width (σ^2) of the RBF kernel, to optimum values. The optimization inside the grid search was driven by TFO crossvalidation. Prior to SVM training all descriptors were normalized in a range [−1, 1]. The toolbox used to implement the SVM with RBF kernel (RBF-SVM) was LIBSVM for Matlab by Chang and Lin,[52] which was downloaded from http://www.csie.ntu.edu.tw/cjlin/libsvm/.

2.5. Validation of Model. The efficiency of the SVM predictor for the classification problem was evaluated using a set of statistics listed below.

The overall accuracy is

$$Q2 = \frac{p}{N} \quad (6)$$

where p is the total number of correctly predicted inhibition complexes and N is the total number of inhibition complexes.

The Mathew's correlation coefficient Cr is defined as

$$Cr(s) = \frac{[p(s)n(s) - u(s)o(s)]}{D} \quad (7)$$

where D is the normalization factor

$$D = [(p(s) + u(s))(p(s) + o(s))(n(s) + u(s))(n(s) + o(s))]^{1/2} \quad (8)$$

for each class s (+ and − for stable and unstable inhibition complexes), $p(s)$ and $n(s)$ are the number of correct predictions and correctly rejected assignments, respectively and $u(s)$ and $o(s)$ are the number of under- and overpredictions.

The coverage for each discriminant structure s is evaluated as

$$Q_S = \frac{p(s)}{p(s) + u(s)} \quad (9)$$

where $p(s)$ and $u(s)$ are the same as in eq 10

The accuracy for s is computed as

$$P_S = \frac{p(s)}{p(s) + o(s)} \quad (10)$$

where $p(s)$ and $o(s)$ are the same as in eq 8.

$Q(+)$ and $Q(-)$ are sensitivity and specificity of stable class prediction, while $P(+)$ and $P(-)$ are precision scores.

Optimum models were selected with maximum F-score known as the harmonic mean of sensitivity and positive precision given as follows:

$$F = 2 \times \frac{(Q(+) \times P(+))}{(Q(+) + P(+))} \quad (11)$$

3. RESULTS

3.1. Clustering Analysis. Selectivity of the 19 kinase families was first explored by cluster analysis. Two sets of variables, structural fragments and 2D autocorrelation vectors were calculated for the 1200 active ligands on the data set. Afterward, *k*-means clustering algorithm yielded 19 ligand clusters (same as the number of kinase families) for each descriptor type. To evaluate the efficiency of kinase selectivity in the two descriptor's sets, we matched ligand clusters with the different kinase families.

Distributions of the two sets of 19 ligand clusters on the 19 kinase families are shown in Figure 1. The cluster distributions depict different patterns. Tyrosine kinase family was the most populated in both graphs but distributions of the ligand clusters on this kinase family differ. In addition, we built two dendrograms by clustering kinase families according to the distributions of the ligand clusters in the families. Figure 2 depicts dendrograms for both descriptor types showing 5 clusters. The dendrogram in Figure 2A exhibits similar distributions of kinase families in the clusters in comparison with the dendrogram in Figure 2B. The five clusters in Figure 2A represent 4 kinase families and a group of 15 kinase families. The single-family clusters are tyrosine kinase, cyclin, AGC Ser/Thr protein kinase, and atypical kinases. In turn, Figure 2B depicts three single-family clusters, tyrosine kinases, AGC Ser/Thr protein kinases, and atypical kinases, one cluster with two families, CMGC Ser/Thr protein kinase and TKL Ser/Thr protein kinase, and a ten-family cluster. The distributions of the kinase families in the several ten-families clusters differ at lower squared Euclidean distances.

The clustering points out that independently allocated kinase families are inhibited by different chemical scaffolds. At the same time, families in multiple clusters could share some active ligand similarity depending on each encoding scheme. Especially, the allocation of CMGC Ser/Thr protein kinase and TKL Ser/Thr protein kinase in the same cluster in Figure 2B suggests a similar inhibition scaffold for these families according to the topological approach.

We also evaluated the ability of a fragment and its topological descriptors to distinguish between active and inactive chemical scaffolds in the data set. The mean squared differences between the normalized descriptors of active and inactive inhibitors were 0.361 and 0.334 from the fragment- and topological-based approaches, respectively. These values significantly differed from the mean squared differences of 100 scrambled partitions of active and inactive ligands, which were about 10^{-2} for both descriptor types. Topological autocorrelation descriptors differentiate slightly better between active and inactive inhibitors in comparison with the fragment approach.

Besides this preliminary qualitative analysis, we attempted to develop classifiers of the ligand affinity toward kinases by SVM training. In silico modeling of the kinomics can be performed using PCM because it offers an interesting framework for multitarget inhibition problems. Interactions between a series of proteins and a series of ligands can be considered in a unique model. Besides predicting the affinity of new ligands toward a group of related targets, prediction of affinities for the new proteins against a set of ligands can also be calculated. In this way, inhibition complexes rather than isolated inhibitors or targets are considered as a unified data set for model development. In our study, structural fragments and topological autocorrelation features of targets and ligands were combined to train the SVM classifiers with the stability of kinase-inhibitor complexes (experimental values of IC_{50}) forming the target vectors. Kinase−inhibitor complexes were classified as low-affinity (unstable) and high-affinity (stable) inhibition complexes.

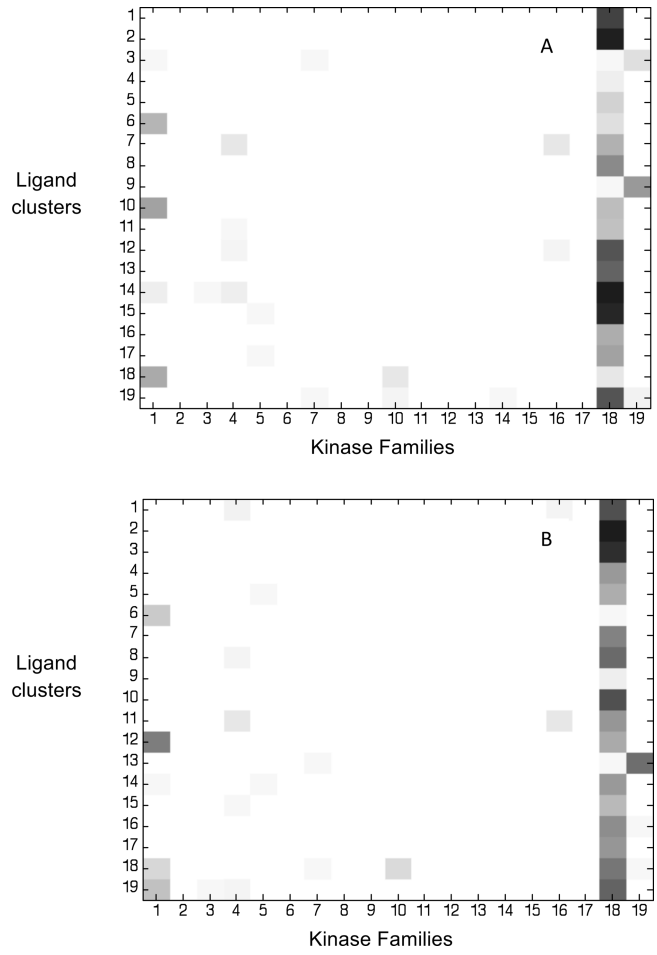


Figure 1. Density map representing occurrence ratios of kinase families on active ligand clusters obtained by atom-center fragment descriptors (A) and 2D autocorrelation vectors (B). Kinase families: (1) AGC Ser/Thr protein kinase, (2) CAMK Ser/Thr protein kinase, (3) CAMP-dependent kinase regulatory chain, (4) CMGC Ser/Thr protein kinase, (5) cyclin, (6) DCK/DGK, (7) herpes virus thymidine kinase, (8) PDGF/VEGF growth factor, (9) PI3/PI4-kinase, (10) PI3K p85 subunit, (11) PP1 inhibitor, (12) phosphoglycerate kinase, (13) phosphorylase b kinase regulatory chain, (14) STE Ser/Thr protein kinase, (15) Ser/Thr protein kinase, (16) TKL Ser/Thr protein kinase, (17) thymidine kinase, (18) Tyr protein kinase, and (19) atypical kinase.

3.2. Structural Fragments Classifier. SF and SSF descriptors computed for the kinase sequences and the 2D structural sketches of the inhibitors were combined into a single feature matrix by simple concatenation of target and ligand descriptors blocks. The data set was separated into a training set with 2696 inhibition complexes (80%) and a test set with 899 inhibition complexes (20%).

In a first attempt, we implemented a linear kernel but the highest crossvalidation accuracy was only about 65%. Then,

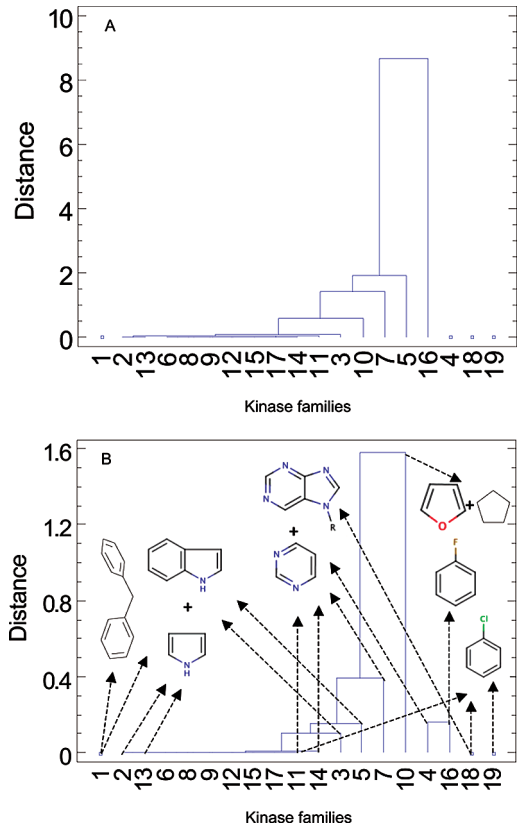


Figure 2. Dendrograms of 5 clusters of kinase families according to occurrences ratios on active ligand clusters obtained by fragment descriptors (A) and 2D autocorrelation vectors (B). Kinase families: (1) AGC Ser/Thr protein kinase, (2) CAMK Ser/Thr protein kinase, (3) CAMP-dependent kinase regulatory chain, (4) CMGC Ser/Thr protein kinase, (5) cyclin, (6) DCK/DGK, (7) herpes virus thymidine kinase, (8) PDGF/VEGF growth factor, (9) PI3/PI4-kinase, (10) PI3K p85 subunit, (11) PP1 inhibitor, (12) phosphoglycerate kinase, (13) phosphorylase b kinase regulatory chain, (14) STE Ser/Thr protein kinase, (15) Ser/Thr protein kinase, (16) TKL Ser/Thr protein kinase, (17) thymidine kinase, (18) Tyr protein kinase, (19) atypical kinase.

Table 1. Crossvalidation and Prediction Statistics for the Training and Test Sets According to the Fragment SVM Model for the Classification of the Stability of Kinase Inhibition Complexes[a]

experiment	$Q2$	$Q(+)$	$Q(-)$	$P(+)$	$P(-)$	Cr
training set crossvalidation	0.78	0.75	0.79	0.65	0.86	0.53
test set prediction	0.78	0.77	0.77	0.65	0.88	0.55

[a] + and −: the indexes were evaluated for stable (IC_{50} < 1 μM) and unstable (IC_{50} > 1 μM) kinase inhibition complexes, respectively. $Q2$ is the number of correct predictions/number of examples; $P(s)$ is the number of correct prediction for class s/all prediction made for s; $Q(s)$ is the number of correct prediction for class *s*/observed in class *s*; Cr is Matthews's correlation coefficient. $Q(+)$ and $Q(-)$ are sensitivity and specificity of stable class prediction, and $P(+)$ and $P(-)$ are precision scores.

adjusting SVM parameters throughout a TFO crossvalidation in a grid search yielded a nonlinear classifier with crossvalidation results shown in Table 1. An overall TFO crossvalidation accuracy of 78% for the classification of inhibition complexes was achieved with a correlation coefficient $Cr = 0.53$. It is noteworthy that crossvalidation accuracies for recognizing stable $Q(+) = 0.75$ and unstable inhibition complexes $Q(-) = 0.78$ are equivalent to the overall accuracy achieved. Taking into account that the predictor was trained with structural fragment information from targets and ligands, these accuracies about 80% for recognizing stable and unstable inhibition complexes are adequate.

Table 2. Crossvalidation and Prediction Statistics for the Training and Test Sets According to the Optimum Topological SVM Model with for the Classification of the Stability of Kinase Inhibition Complexes[a]

experiment		$Q2$	$Q(+)$	$Q(-)$	$P(+)$	$P(-)$	Cr
training set crossvalidation	complex-wise	0.82	0.85	0.81	0.69	0.92	0.63
	ligand-wise	0.82	0.84	0.81	0.68	0.91	0.62
	kinase-wise	0.65	0.75	0.60	0.48	0.83	0.32
test set prediction	complex-wise	0.81	0.87	0.78	0.67	0.92	0.62
	ligand-wise	0.82	0.91	0.77	0.69	0.94	0.66

[a] SVM parameters were $\sigma^2 = 0.091$ and $C = 1.36$. + and -: The indexes were evaluated for stable (IC_{50} < 1 μM) and unstable (IC_{50} > 1 μM) kinase inhibition complexes, respectively. $Q2$ is the number of correct predictions/number of examples; $P(s)$ is the number of correct prediction for class *s*/all prediction made for *s*; $Q(s)$ is the number of correct prediction for class *s*/observed in class *s*; Cr is Matthews's correlation coefficient. $Q(+)$ and $Q(-)$ are sensitivity and specificity of stable class prediction, and $P(+)$ and $P(-)$ are precision scores.

3.3. Topological Autocorrelation Classifier. Topological features were computed from the primary sequence of the kinases and the 2D sketch of the inhibitor structure descriptors (Table 2SI in Supporting Information). Similar to the fragment-based classifier, the data set was separated into training and test sets in the same way.

A linear kernel could only produce a poor performance (67% accuracy) in a cross-validated training. Then, adjusting SVM parameters throughout a TFO crossvalidation in a grid search yielded an optimum nonlinear SVM classifier with crossvalidation results shown in Table 2. In Table 2, complex-wise statistics refers to the basic crossvalidation experiments in which only nonredundant complexes were included. In this case, an overall complex-wise TFO crossvalidation accuracy of 82% was achieved for the classification of inhibition complexes with a correlation coefficient $Cr = 0.63$. Crossvalidation accuracies to identify stable $Q(+) = 0.85$ and unstable inhibition complexes $Q(-) = 0.81$ resulted similar to the overall accuracy. These accuracies and the correlation coefficient are higher than the statistics reported in Table 1 for the fragment-based predictor.

Fragment descriptors are more intuitive and easy to interpret, but they only account for substructure occurrences on the structure and lack the information of connectivity and sequence order. In contrast, 2D autocorrelation vectors account for property distributions on the topological structure accounting for atom arrangements in the bidimensional molecular sketch and amino-acid residue distributions along the protein sequence. In view of the training set results, we conclude that autocorrelation approach outperforms the fragment-based classifier and it is more convenient for modeling kinase inhibition.

In addition, the stability of the optimum topological model to recognize new ligands and kinases was evaluated. We performed two additional crossvalidation experiments in which kinase complex sharing similar ligands were kept in the same data subset during crossvalidation, we called this experiments ligand-wise crossvalidation, similarly we run another crossvalidation in which complexes of the same kinase were kept in the same subset and this was called

Table 3. Ligand-Wise TFO Crossvalidation Accuracies for the 19 Kinase Families in the Training Set According to the Optimum Topological SVM Model[a]

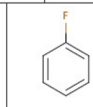
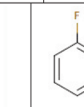
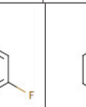
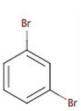
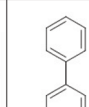
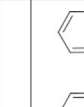
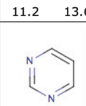
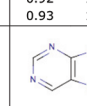
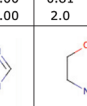
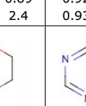
kinase family	*Q2*	*Q*(+)	*Q*(−)	occurrence ratio (%)
AGC Ser/Thr protein kinase	0.80	0.85	0.76	7.97
CAMK Ser/Thr protein kinase	0.25	0.00	0.33	0.15
CAMP-dependent kinase regulatory chain	0.96	0.00	1.00	3.52
CMGC Ser/Thr protein kinase	0.74	0.85	0.54	1.45
cyclin	0.75	0.00	0.95	1.04
DCK/DGK	1.00		1.00	0.41
herpes virus thymidine kinase	0.87	0.14	0.98	1.93
PDGF/VEGF growth factor				0.00
PI3/PI4-kinase	1.00		1.00	1.48
PI3K p85 subunit	1.00	1.00		0.37
PP1 inhibitor	0.98	0.00	1.00	1.63
Phosphoglycerate kinase	1.00		1.00	4.15
Phosphorylase b kinase regulatory chain	0.67	0.00	1.00	0.11
STE Ser/Thr protein kinase	0.50	0.00	1.00	0.07
Ser/Thr protein kinase	1.00		1.00	0.22
TKL Ser/Thr protein kinase	0.77	1.00	0.44	0.82
thymidine kinase	1.00		1.00	0.67
Tyr protein kinase	0.80	0.87	0.76	68.81
atypical kinase	0.76	0.75	0.76	5.19

[a] + and −: the indexes were evaluated for stable ($IC_{50} < 1$ μM) and unstable ($IC_{50} > 1$ μM) kinase inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *Q*(*s*) is the number of correct prediction for class *s*/observed in class *s*. *Q*(+) and *Q*(−) are sensitivity and specificity of stable class prediction.

kinase-wise crossvalidation. The results of these experiments are reported in Table 2, overall accuracy for ligand-wise crossvalidation was 82%, illustrating that the model correctly classifies the affinity of novel ligands toward existing kinases. All further reported statistical analysis for kinase families (Table 3 and 5) and ligand chemotypes (Table 4) were performed on ligand-wise crossvalidation. Interestingly, the kinase-wise 10-fold-out crossvalidation in Table 2 showed that the model differentiated complexes of new kinase with overall accuracy about 65% and stable complexes with accuracy of 75%. This result, although discrete, is noteworthy taking into account that when removing highly represented kinases from the training subset also large series of inhibitors are left out. This fact corroborates the relevance of the topological feature space to model kinase inhibition as well as the self-consistency of the optimum SVM model.

3.4. Performance of the Optimum Topological Autocorrelation Classifier for Different Kinase Families and Chemotypes. Kinase inhibition data set includes inhibitory activities of a diverse chemical space toward 19 kinase families. It is very interesting to analyze the optimum classifier performance for each kinase family in the data set. The classification accuracies of SVM predictor for each kinase family are shown in Table 3. The predictor performance was very homogeneous to all families. The overall accuracies for the recognition of stable and unstable inhibition complexes were higher than 67% for all but one the kinase families. However, the classifier was unable to recognize stable inhibition complexes of seven protein kinase families with low occurrences of stable complexes in the crossvalidation experiments. This fact suggests that the training set information is very diversified and generalization from one family to the others is very difficult inside the training set. In addition to the low statistical significances of the stable inhibitors of these families in the data set, another factor accounting for these low accuracies in crossvalidation experiments could be the complexity of the target-ligand interactions for these kinase families. In this regard, a recent review on QSAR modeling of binding affinities stated that conformational changes and binding site flexibility lead to the conclusion

Table 4. Ligand-Wise TFO Crossvalidation and Prediction Accuracies for Training and Test Sets for 30 Substructures in the Kinase Inhibitors Data Set According to the Optimum Topological Classifier[a]

Substr.										
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q2	0.83	0.81	1.00	1.00	0.81	0.94	0.80	0.90	0.84	0.86
Q(+)	0.83	0.85	1.00	1.00	0.77	0.92	0.80		0.75	1.00
Q(-)	0.83	0.79			0.86	1.00	0.79	0.90	0.85	0.83
Occ.(%)	14.0	13.2	0.4	0.7	1.3	1.9	1.6	1.1	2.3	3.1
Substr.										
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q2	0.81	0.80	0.84	0.85	0.73	0.72	0.79	0.74	0.89	1.00
Q(+)	0.84	0.87	0.95	0.91	0.92	0.92	0.80	0.82	1.00	1.00
Q(-)	0.79	0.76	0.81	0.80	0.50	0.50	0.79	0.70	0.83	1.00
Occ.(%)	91.8	92.5	3.1	2.9	26.1	25.3	19.3	22.8	0.7	1.3
Substr.										
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q2	0.86	0.75	0.94	1.00	0.79	0.80	0.81	0.82	0.92	1.00
Q(+)	0.74	0.92	1.00	1.00	0.78	0.85	0.45	0.88	1.00	1.00
Q(-)	0.97	0.60	0.93	1.00	0.80	0.79	0.90	0.80	0.91	1.00
Occ.(%)	2.7	3.1	1.1	0.8	56.8	59.3	3.8	3.7	2.3	2.3
Substr.										
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q2	0.77	0.77	0.69	0.76	0.86	0.84	0.80	0.50	0.86	0.89
Q(+)	0.89	0.94	0.97	0.96	0.90	1.00	1.00	1.00	0.95	0.96
Q(-)	0.66	0.62	0.30	0.40	0.81	0.71	0.67	0.00	0.67	0.65
Occ.(%)	22.3	17.9	7.3	7.9	5.7	4.1	0.2	0.2	10.8	10.9
Substr.										
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q2		1.00	0.78	0.75	0.93	1.00	0.74	0.77	0.93	1.00
Q(+)		1.00	0.78	0.78	1.00	1.00	0.68	0.69	1.00	1.00
Q(-)			0.79	0.73	0.92	1.00	0.81	0.89	0.92	1.00
Occ.(%)	0.0	0.1	11.2	13.6	0.93	1.00	2.0	2.4	0.93	1.00
Substr.										
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test
Q2	0.96	0.93	0.84	0.83	0.95	0.99	0.89	0.95	0.80	0.83
Q(+)	0.70	0.71	0.91	0.92	0.38	1.00	0.97	1.00	0.91	0.91
Q(-)	0.98	0.97	0.77	0.75	0.98	0.99	0.80	0.83	0.56	0.56
Occ.(%)	5.5	4.9	36.5	36.4	7.8	8.2	2.3	2.1	7.9	8.0

[a] + and −: the indexes were evaluated for "stable" ($IC_{50} < 1$ μM) and "unstable" ($IC_{50} > 1$ μM) kinase inhibition complexes, respectively. *Q2* is the number of correct predictions/number of examples; *Q*(*s*) is the number of correct prediction for class s/observed in class s. *Q*(+) and *Q*(−) are sensitivity and specificity of stable class prediction. Occ.(%) is the occurrence ratio.

that similar analogs bind to the same binding site in different modes. Furthermore, the binding site residues in the ligand-protein interactions are not the same due to the difference in the flexible binding site residues.[53]

Although our approach is alignment- and conformation-independent, identical or closely related inhibitor structures, which interact in different ways, can cause model failure for such inhibition complexes. However, one of the advantages of the alignment- and structure-free protein activity/function prediction methods is that they are less prone to be affected by protein folding or ligand's binding orientation. At the same time, for QSAR studies, when the binding mechanism and orientation are unknown and if a broad variety of targets are processed, it is usually accepted that more robust and accurate models can be derived from 2D-structure encoding frameworks rather than 3D detailed description of the molecules.

According to Table 3, kinase families with the highest occurrence ratios on the training set exhibit crossvalidation accuracies higher than 80%. Those kinase families are Tyr protein kinase, AGC Ser/Thr protein kinase, Phosphoglycerate kinase and CAMP-dependent kinase regulatory chain families with occurrence ratios of 68.81%, 7.97%, 4.15%, and 3.52%, respectively. Among these families, tyrosine protein kinase has been most studied for targeting cancer. Tyrosine protein kinase directly participates in cell growth through the signal passing pathways. Five types of proteins participate in the growth control of mammalian cells: growth factors, growth factor receptors, intracellular transducers, nuclear transcription factors, and cell cycle control proteins. Some cell surface receptors have an extracellular ligand-binding domain attached to an integral protein tyrosine kinase in their cytoplasmic domain. These receptors transmit the growth signal by phosphorylating their tyrosine residues as well as one or more target of proteins, thus initiating a cascade of events.[21] It is widely accepted that protein tyrosine kinases play a fundamental role in cancer. That is why they became attractive therapeutic targets and it has provided impetus for an extensive effort to develop specific inhibitors of these enzymes as chemotherapeutic agents. An overall high performance (~80%) of our classifier in predicting the stable versus unstable complexes in this kinase family, therefore, has very useful practical implications to the inhibitor design problem.

We also analyzed the predictor behavior for different chemical subspaces on the inhibitor data set. In this sense, 30 substructural templates were considered for comparing the classifier accuracy regarding the different chemotypes on the modeled chemical space. As can be observed in Table 4, all the analyzed substructures showed overall accuracies about or higher than 70% and accuracies for separate classes were lower than 50% only for low affinity ligands bearing 1,3-dichlorobenzene substructure and high affinity ligands bearing *m*-methyltoluene. From this result we conclude that the classifier performed well over the chemical space represented by the kinase inhibitors in the training set.

3.5. Prediction of the Test Set. Crossvalidation accuracy gives an estimate of the internal consistency of the predictive models but a more realistic measurement of the prediction power can be achieved by predicting a blind test set. In Table 2, we show that the results from the topological-based classifier on such data sets also perform well and lead to an overall accuracy of about 81% for the complex-wise evaluation. This is promising considering the fact that the test set prediction accuracies were in the same range as obtained in crossvalidation experiments of the training set in Table 2, thus excluding the possibility of overfitting. Classification of test set also performed well for the topological models taking into account that test set accuracy of the fragment-based predictor was 79% (Table 1).

Table 5. Ligand-Wise Prediction Accuracies for the 19 Kinase Families in the Kinase Inhibitor in the Test Set According to the Optimum SVM Model[a]

kinase family	*Q*2	*Q*(+)	*Q*(−)	occurrence ratio (%)
AGC Ser/Thr protein kinase	0.70	0.72	0.68	8.57
CAMK Ser/Thr protein kinase	1.00		1.00	0.33
CAMP-dependent kinase regulatory chain	0.97	0.00	1.00	4.23
CMGC Ser/Thr protein kinase	1.00	1.00	1.00	0.67
cyclin	1.00	1.00	1.00	1.22
DCK/DGK	1.00		1.00	0.67
herpes virus thymidine kinase	0.86		0.86	0.78
PDGF/VEGF growth factor	1.00		1.00	0.11
PI3/PI4-kinase	1.00		1.00	1.45
PI3K p85 subunit	1.00	1.00		0.67
PP1 inhibitor	1.00		1.00	0.78
phosphoglycerate kinase	1.00		1.00	4.56
phosphorylase b kinase regulatory chain				0.00
STE Ser/Thr protein kinase	1.00	1.00	1.00	0.22
Ser/Thr protein kinase	1.00		1.00	0.33
TKL Ser/Thr protein kinase	0.57	0.67	0.50	0.78
thymidine kinase	1.00		1.00	0.56
Tyr protein kinase	0.79	0.89	0.72	69.52
atypical kinase	0.78	0.85	0.75	4.56

[a] + and −: The indexes were evaluated for stable ($IC_{50} < 1$ μM) and unstable ($IC_{50} > 1$ μM) kinase inhibition complexes, respectively. *Q*2 is the number of correct predictions/number of examples; *Q*(*s*) is the number of correct prediction for class *s*/observed in class *s*. *Q*(+) and *Q*(−) are sensitivity and specificity of stable class prediction.

In addition, the ability of the topological predictor to recognize totally new ligands was estimated by evaluating the model accuracy for 462 out of the 899 kinase inhibition complexes in the test set for which ligands information was not available in the training set. This predictor correctly classified 82% of the 462 kinase complexes with totally new ligands. Remarkably, stable complexes were recognized with accuracy of 91% while the accuracy for the unstable complexes was 77%. This result showed that the optimum classifier not only correctly learned the kinase inhibition pattern, but that the learned pattern was adequately generalized to the test set, including totally new high-affinity ligands.

Table 5 shows classification results for each kinase family on the test set. All kinase families in the test were classified with overall accuracies >55%. Furthermore, classifier performance has accuracies over 80% for 16 out of the 18 kinase families in the test set. The model failed to recognize only high affinity ligands of CAMP-dependent kinase regulatory chain family. Thus, the information from the training set was successfully generalized and the prediction results were very homogeneous to the majority of the kinase families in the test set.

Similarly, the analysis of the classifier performance on the test set according to different chemotypes shows that the

classifier attained similar performance to the training sets used in crossvalidation in Table 4. All the inhibitor types have overall accuracies about or higher 50% and only 1,3-difluorobenzene has accuracy <70%. Low affinity ligands bearing 1,3-diclorobenzene chemotypes were classified with low accuracies and the predictor failed to recognized low affinity 1,3-difluorobenzene derivatives. Despite of the low prediction performance for these chemotypes, the overall performance of the predictor on this blind set is adequate. The classifier recognized the inhibition pattern from different kinase families and also properly discriminated between stable and unstable inhibitor complexes belonging to several chemical subspaces in the test set.

The differential relevance of the topological autocorrelation space for modeling kinase inhibition was evaluated by sensitivity analysis.[36] The impact of each variable in the optimum model was estimated by measuring test set prediction accuracies for different modified feature matrices in which each autocorrelation vector at a time was replaced by a constant vector of same length. The magnitude of the importance of each input variable in the predictor was taken as the underperformance score defined as the ratio between *F*-scores (eq 11) for original and modified feature matrices. The sensibility analysis yields top-9 relevant autocorrelation vectors in descending order as follows: AASA3H_t > AASA2R_a > *A*ASA4R_a > AASA5R_a > AASA4ASA$_N$ > AASA5ASA$_N$ > ATS6v > ATS3e > ATS4e. It is noteworthy that the most relevant inputs are kinase's autocorrelations of hydrophobicity/polarity-related properties such as thermodynamic transfer hydrophobicity (H_t), solvent-accessible reduction ratio (R_a) and solvent-accessibility area for native state (ASA$_N$), in combination with ligand's autocorrelations of atomic volumes and electronegativities on the 2D structure sketch.

4. DISCUSSION

Inhibition of protein kinases can be broadly classified into three categories: ATP-competitive inhibition, substrate-competitive inhibition, and allosteric inhibition. Successful treatments of chronic myeloid leukemia and gastrointestinal stromal tumor with Gleevec[8] have recently drawn much attention because of its excellent selectivity and its ability to bind to a precise inactive conformation of Abl kinase. However, the emergence of drug-resistant mutants[54] and structural studies suggest that mutations in the kinase domain cause resistance to the Abl kinase inhibitor.[55] Other studies have also shown that some inhibitors can recognize specific inactive conformation of B-Raf (1UWH) and p38 (1W83), whereas some others can inhibit the active form of Abl kinase.[56] However, all of them have been shown to decrease or completely lose inhibitory activity toward some mutated kinase. In this sense, Thaimatta et al.[9] in the review of kinase inhibitors stated that the modulation of kinase activity has not been sufficiently exploited for therapeutic purposes. These authors suggested that inhibition of a single kinase may be insufficient to achieve a therapeutic benefit, and that promiscuous small-molecule kinase inhibitors or cocktails of inhibitors may be more promising than selective agents by targeting several kinases. In view of these facts, different computational approaches for kinase drug design need to be exploited to find novel, more efficient and side-effect-free kinase inhibitors.

QSAR and docking techniques have formed the basis for predicting binding affinities in most cases.[57] However, even if protein targets were closely related or if they correspond to the same protein family, different targets should constitute corresponding training sets to model multiple target-ligand systems. In this sense, when modeling affinities toward multiple targets; this technique produces a huge amount of models, each one applicable to a target. This makes the model's interpretation and generalization difficult even for closely related targets. On the other hand, docking studies employ 3D structures from targets and ligands for generating interaction scores for ligand-target binding conformations. The main uses of docking are to select hits in virtual library screening and to evaluate 3D binding modes. This technique is very convenient to preselect "true binders" to build a chemical library from a virtual library. A detailed description of target and ligand 3D structures is needed for docking. However, generalizations can be made only for closely related targets for which very good alignments are available and ligands are also very similar.[53] Generalized models for kinase inhibition that include ligands from different scaffolds will be difficult to build by docking.

A large number of QSAR models on kinase inhibition have been reported and tyrosine kinases are among the most widely studied families. In this regard, Kurup et al.[20] published a review of QSAR studies for the inhibitory activity of a very varied chemical data set toward five tyrosine kinases: epidermal growth factor receptor tyrosine kinase, platelet-derived growth factor receptor tyrosine kinase, fibroblast growth factor receptor tyrosine kinase of vascular endothelial growth factor receptor tyrosine kinase, and non receptor tyrosine kinase. They reported a huge amount of 40 QSAR equations using hydrophobicity, steric, and electronic descriptors. The authors did not use target information but they tried to establish target-ligand interaction hypothesis by comparing quality and descriptor occurrences on the models for different inhibitor data sets on the same target or the same inhibitor data set for different targets. Despite that rigor in this area, the authors used very intuitive descriptors, and although the stability of target-ligands were predicted with high crossvalidation accuracies, the use and generalization of 40 models, as well as their comparative interpretation is rather rough.

Kinase sequence information had also been correlated with inhibition selectivity. A novel approach combines the understanding of small molecules and target sequence and genes and, thereby, assists researchers in finding new targets for existing molecules or understanding selectivity and polypharmacology of molecules in related targets. Chemogenomics combines genomic data, structural biological data, classical dendrograms, and selectivity data to explore, define, and classify the medicinally relevant target space for any relevant biological system. Consequently, exploitation of this information in the discovery of kinase inhibitors defines practical kinase chemogenomics (kinomics).[58] The authors presented the first dendrogram of kinases based entirely on small molecule selectivity data. They found that the selectivity dendrogram varied from sequence-based clustering due to the higher-level groupings of the smallest clusters, and it remains very comparable for closely homologous targets. As

a main result, it was found that the smaller comparable molecules inhibit higher homologous kinases in a more desirable way.

In our study, we also employed ligand clustering to evaluate and to differentiate the ability of ligand cluster distributions in kinase families. Furthermore, optimum modeling of protein−ligand interactions was developed combining topological descriptors of targets and ligands. Autocorrelation vectors weighted by amino acids/residues and atomic properties encode target sequences and inhibitor structures. Those descriptors account for amino-acid distributions on the target sequences and atom distribution on the 2D sketch of the inhibitor molecules. Furthermore, while combining topological autocorrelation features, interactions between target and ligand structures are encoded in a conformation-independent set of descriptors. The most relevant autocorrelation features were found to be thermodynamic transfer hydrophobicity (H_t), solvent-accessible reduction ratio (R_a) and solvent-accessibility area for native state ASA_N, which encode a kinases inhibitory pattern, defined by the distributions of hydrophobicity/polarity states along the sequence. At the same time, the differential affinity of ligands toward kinases was ruled by the distributions of atomic volume and electronegativity on the 2D structure sketches. To the best of our knowledge, our study is the first model for predicting inhibition data on 62 kinases and a wide chemical space, which allows discriminating between stable and unstable inhibition complexes with adequate accuracies about 82% for training set crossvalidation and test sets. The predictor is available online at http://gibk21.bse.kyutech.ac.jp/AUTOkinI/SVMpredictor.html.

5. CONCLUSIONS

Classical QSAR studies are mainly ligand-based approaches whereas PCM also consider target structural information for predicting binding affinities. SVMs trained with PCs extracted from the structural fragment interaction and topological autocorrelation interaction matrices could classify the stability of a large data set of kinase inhibition complexes. We applied the fragment and topological approximations to the proteochemometric modeling for the inhibition of 62 kinases from 19 protein families. The topological model was superior to the fragment-based classifier with maximum crossvalidation accuracies about 82% for training set crossvalidation and test set prediction. Furthermore, test set accuracies of the optimum topological classifier were very homogeneous across different kinase families and substructural fragments of the ligands. The present method can be applied to other protein−ligand interactions.

ACKNOWLEDGMENT

This work has been supported in part by Grants-in-Aid for Scientific Research 20016022, 21310131 to A.S. from Ministry of Education, Culture, Sports, Science and Technology in Japan.

Supporting Information Available: Excel tables with names and values of the amino acids/residues properties used to calculated Amino Acid Sequence Autocorrelation (AASA) vectors and autocorrelation descriptor symbols and values and a structure-data file (txt) containing ligand structures, target sequences, activity classes, and training or test set locations in the data set. This information is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Settleman, J. Mutated kinases as targets for cancer drugs. *Drug Discovery Today: Dis. Mech.* **2005**, *2*, 139–144.
(2) Manning, D. B.; Whyte, R.; Martinez, T.; Hunter, T.; Sudarsanam, S. The protein kinase complement of the humangenome. *Science* **2002**, *298*, 1912–1934.
(3) Cohen, P. Protein kinases-the major drug targets of the twenty-first century. *Nat. Rev. Drug Discovery* **2002**, *1*, 309–315.
(4) Faivre, S.; Djelloul, S.; Raymond, E. New paradigms in anticancer therapy: Targeting multiple signaling pathways with kinase inhibitors. *Semin. Oncol.* **2006**, *33*, 407–420.
(5) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243–257.
(6) Fedorov, M.; Sundström, B.; Marsden, B.; Knapp, S. Insights for the development of specific kinase inhibitors by targeted structural genomics. *Drug Discovery Today* **2007**, *12*, 365–362.
(7) Fischer, P. M. The design of drug candidate molecules as selective inhibitors of therapeutically relevant protein kinases. *Curr. Med. Chem.* **2004**, *11*, 1563–1583.
(8) Bogoyevitch, M. A.; Fairlie, D. P. A new paradigm for protein kinase inhibition: blocking phosphorylation without directly targeting ATP binding. *Drug Discovery Today* **2007**, *12*, 622–623.
(9) Thaimattam, R.; Banerjee, R.; Miglani, R.; Iqbal, J. Protein kinase inhibitors: Structural insights into selectivity. *Curr. Pharm. Des.* **2007**, *13*, 2751–2765.
(10) Cavasotto, C. N.; Abagyan, R. A. Protein flexibility in ligand docking and virtual screening to protein kinases. *J. Mol. Biol.* **2004**, *337*, 209–225.
(11) Fousteris, M. A.; Papakyriakou, A.; Koutsourea, A.; Manioudaki, M.; Lampropoulou, E.; Otyepka, M.; Kryŝtof, V.; Havlíĉek, L.; Siglerova, V.; Strnad, M.; Koĉa, J. Docking-based development of purine-like inhibitors of cyclin-dependent kinase-2. *J. Med. Chem.* **2000**, *43*, 2506–2513.
(12) Jing-Fa, X.; Ze-Sheng, L.; Miao, S.; Yuan, Z.; Chia-Chung, S. Homology modeling and molecular dynamics study of GSK3/SHAGGY-like kinase. *Comput. Biol. Chem.* **2004**, *28*, 179–188.
(13) Sheinerman, F. B.; Giraud, E.; Laoui, A. High Affinity targets of protein kinase inhibitors have similar residues at the positions energetically important for binding. *J. Mol. Biol.* **2005**, *352*, 1134–1156.
(14) Miao, S.; Zesheng, L.; Yuan, Z.; Qingchuan, Z.; Chia-chung, S. Homology modeling and docking study of cyclin-dependent kinase (CDK) 10. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2851–2856.
(15) Rockey, W. M.; Elcock, A. H. Structure selection for protein kinase docking and virtual screening: homology models or crystal structures. *Curr. Prot. Pept Sci.* **2006**, *7*, 437–57.
(16) Manetti, F.; Locatelli, G. A.; Maga, G.; Schenone, S.; Modugno, M.; Forli, S.; Corelli, F.; Botta, M. A Combination of Docking/Dynamics Simulations and Pharmacophoric Modeling To Discover New Dual c-Src/Abl Kinase Inhibitors. *J. Med. Chem.* **2006**, *49*, 3278–3286.
(17) Kulkarni, R. G.; Srivani, P.; Achaiah, G.; Sastry, G. N. Strategies to design pyrazolyl urea derivatives for p38 kinase inhibition: a molecular modeling study. *J. Comput.-Aided Mol. Des.* **2007**, *21*, 155–166.
(18) Papadimitriou, E.; Spyroulias, G. A.; Nikolaropoulos, S. S. Pyrrolo[2,3-*a*]carbazoles as potential cyclin dependent kinase 1 (CDK1) inhibitors. Synthesis, biological evaluation, and binding mode through docking simulations. *J. Med. Chem.* **2008**, *51*, 1048–1052.
(19) Ravindra, G. K.; Achaiah, G.; Sastry, G. N. Molecular modeling studies of phenoxypyrimidinyl imidazoles as p38 kinase inhibitors using QSAR and docking. *Eur. J. Med. Chem.* **2008**, *43*, 830–838.
(20) Kurup, A.; Garg, R.; Hansch, C. Comparative QSAR study of tyrosine kinase inhibitors. *Chem. Rev.* **2001**, *101*, 2573–2600.
(21) Woolfrey, J. R.; Weston, G. S. The use of computational methods in the discovery and design of kinase inhibitors. *Curr. Pharm. Des.* **2002**, *8*, 1527–45.
(22) Fernández, M.; Tundidor-Camba, A.; Caballero, J. Modeling of cyclin-dependent kinase inhibition by 1*H*-pyrazolo[3,4-*d*]pyrimidine derivatives using artificial neural networks ensembles. *J. Chem. Inf. Model.* **2005**, *45*, 1884–1895.
(23) González, M. P.; Caballero, J.; Helguera, A. M.; Garriga, M.; González, G.; Fernández, M. 2D autocorrelation modeling of the inhibitory activity of cytokinin-derived cyclindependent kinase inhibitors. *Bull. Math. Biol.* **2006**, *68*, 735–751.

(24) Caballero, J.; Fernandez, M.; Saavedra, M.; Gonzalez-Nilo, F. D. 2D Autocorrelation, CoMFA, and CoMSIA modeling of protein tyrosine kinases' inhibition by substituted pyrido[2,3-*d*]pyrimidine derivatives. *Bioorg. Med. Chem.* **2008**, *16*, 810–821.
(25) Subramanian, J.; Sharma, S.; B-Rao, C. A novel computational analysis of ligand-induced conformational changes in the ATP binding sites of cyclin dependent kinases. *J. Med. Chem.* **2006**, *49*, 5434–5441.
(26) Sperandio da Silva, G. M.; Sant'Anna, C. M. R.; Barreiro, E. J. A novel 3D-QSAR comparative molecular field analysis (CoMFA) model of imidazole and quinazolinone functionalized p38 MAP kinase inhibitors. *Bioorg. Med. Chem.* **2004**, *312*, 159–3166.
(27) Edraki, N.; Hemmateenejad1, B.; Miri, R.; Khoshneviszade, M. QSAR Study of phenoxypyrimidine derivatives as potent inhibitors of p38 kinase using different chemometric tools. *Chem. Biol. Drug. Des.* **2007**, *70*, 530–539.
(28) Wei-min, S.; Qi, S.; Wei, K.; Bao-xian, Y. QSAR analysis of tyrosine kinase inhibitor using modified ant colony optimization and multiple linear regression. *Eur. J. Med. Chem.* **2007**, *42*, 81–86.
(29) Holder, S.; Lilly, M.; Brown, M. L. Comparative molecular field analysis of flavonoid inhibitors of the PIM-1 kinase. *Bioorg. Med. Chem.* **2007**, *15*, 6463–6473.
(30) Cao, H.; Zhang, H.; Zheng, X.; Gao, D. 3D QSAR studies on a series of potent and high selective inhibitors for three kinases of RTK family. *J. Mol. Graphics. Modell.* **2007**, *26*, 236–245.
(31) Duchowicz, P. R.; Castro, E. A. QSAR studies for the pharmacological inhibition of glycogen synthase kinase-3. *Med. Chem.* **2007**, *3*, 393–417.
(32) Singh, S. K.; Dessalew, N.; Bharatam, P. V. 3D-QSAR CoMFA study on oxindole derivatives as cyclin dependent kinase 1 (CDK1) and cyclin dependent kinase 2 (CDK2) inhibitors. *Med. Chem.* **2007**, *3*, 75–84.
(33) González-Díaz, H.; González-Díaz, Y.; Santana, L.; Ubeira, F. M.; Uriarte, E. Proteomics, networks and connectivity indices. *Proteomics.* **2008**, (4), 750–78.
(34) Concu, R.; Dea-Ayuela, M. A.; Perez-Montoto, L. G.; Prado-Prado, F. J.; Uriarte, E.; Bolás-Fernández, F.; Podda, G.; Pazos, A.; Munteanu, C. R.; Ubeira, F. M.; González-Díaz, H. 3D entropy and moments prediction of enzyme classes and experimental-theoretic study of peptide fingerprints in *Leishmania* parasites. *Biochim. Biophys. Acta* **2009**, *12*, 1784–1794.
(35) Concu, R.; Dea-Ayuela, M. A.; Perez-Montoto, L. G.; Bolas-Fernández, F.; Prado-Prado, F. J.; Podda, G.; Uriarte, E.; Ubeira, F. M.; González-Díaz, H. Prediction of enzyme classes from 3D structure: a general model and examples of experimental-theoretic scoring of peptide mass fingerprints of *Leishmania* proteins. *J. Proteome Res.* **2009**, *9*, 4372–4382.
(36) Gonzalez-Díaz, H.; Saiz-Urra, L.; Molina, R.; Santana, L.; Uriarte, E. A model for the recognition of protein kinases based on the entropy of 3D van der Waals interactions. *J Proteome Res.* **2007**, *2*, 904–908.
(37) González-Díaz, H.; Saíz-Urra, L.; Molina, R.; González-Díaz, Y.; Sánchez-González, A. Computational chemistry approach to protein kinase recognition using 3D stochastic van der Waals spectral moments. *J. Comput. Chem.* **2007**, *6*, 1042–1048.
(38) Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. S. Development of proteo-chemometrics: A novel technology of use for analysis of drug-receptor interactions. *Biochem. Biophys. Acta* **2001**, *1525*, 180–190.
(39) Caballero, J.; Fernández, L.; Garriga, M.; Abreu, J. I.; Collina, S.; Fernández, M. Proteometric study of ghrelin receptor function variations upon mutations using amino acid sequence autocorrelation vectors and genetic algorithm-based least square support vector machines. *J. Mol. Graphics. Modell.* **2007**, *26*, 166–178.
(40) Caballero, J.; Fernández, L.; Abreu, J. I.; Fernández, M. Amino acid sequence autocorrelation vectors and ensembles of Bayesian-regularized genetic neural networks for prediction of conformational stability of human lysozyme mutants. *J. Chem. Inf. Model.* **2006**, *46*, 1255–1268.
(41) Fernández, L.; Caballero, J.; Abreu, J. I.; Fernández, M. Amino acid sequence autocorrelation vectors and Bayesian-regularized genetic neural networks for modeling protein conformational stability: Gene V protein mutants. *Proteins* **2007**, *67*, 834–852.
(42) Fernández, M.; Abreu, J. I.; Caballero, J.; Garriga, M.; Fernández, L. Comparative modeling of the conformational stability of chymotrypsin inhibitor 2 protein mutants using amino acid sequence autocorrelation (AASA) and amino acid 3D autocorrelation (AA3DA) vectors and ensembles of Bayesian-regularized genetic neural networks. *Mol. Simulat.* **2007**, *13*, 1045–1056.
(43) Ahmad, S.; Kitajima, K.; Selvaraj, S.; Kubodera, H.; Sunada, S.; An, J.-H.; Sarai, A. Protein−ligand interactions: ProLINT database and QSAR analysis. *Genome Inf.* **2003**, *14*, 537–538.
(44) Wu, C. H.; Apweiler, R.; Bairoch, A.; Natale, D. A.; Barker, W. C.; Boeckmann, B.; Ferro, S.; Gasteiger, E.; Huang, H.; Lopez, R.; Magrane, M.; Martin, M. J.; Mazumder, R.; O'Donovan, C.; Redaschi, N.; Suzek, B. The universal protein resource (UniProt): An expanding universe of protein information. *Nucleic Acids Res.* **2006**, *34*, D187–D191.
(45) *Instant JChem ChemAxon*, version 2.1.1; ChemAxon Ltd.: Budapest, Hungary, 2007.
(46) *DRAGON*, version 3.0; Milano Chemometrics: Milan, Italy, 2003.
(47) *MATLAB*, version 7.0; The Mathworks Inc.: Natick, MA, 2006.
(48) Fernandez, M.; Abreu, J. I. *Protmetrics*; Molecular Modeling Group, University of Matanzas: Matanzas: Cuba, 2006.
(49) (a) Bauknecht, H.; Zell, A.; Bayer, H.; Levi, P.; Wagener, M.; Sadowski, J.; Gasteiger, J. Locating biologically active compounds in medium-sized heterogeneous datasets by topological autocorrelation vectors: Dopamine and benzodiazepine agonists. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1205–1213. (b) Moreau, G.; Broto, P. Autocorrelation of a topological structure: A new molecular descriptor. *Nouv. J. Chim.* **1980**, *4*, 359–360.
(50) (a) Nakai, K.; Kidera, A.; Kanehisa, M. Cluster analysis of amino acid indices for prediction of protein structure and function. *Protein Eng.* **1988**, *2*, 93–100. (b) Tomii, K.; Kanehisa, M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Protein Eng.* **1996**, *9*, 27–36. (c) Kawashima, S.; Kanehisa, M. AAindex: Amino acid index database. *Nucleic Acids. Res.* **2000**, *28*, 374–374.
(51) (a) Cortes, C.; Vapnik, V. Support-vector networks. *Mach. Learn.* **1995**, *20*, 273–297. (b) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. Knowl. Discovery.* **1998**, *2*, 1–47. (c) Vapnik, V. *Statistical Learning Theory*; New York: Wiley; 1998.
(52) Chih-Chung, C. Chih-Jen, L. LIBSVM: A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm (Accessed May 23, 2007).
(53) Kim, K. H. Outliers in SAR and QSAR: 2. Is a flexible binding site a possible source of outliers? *J. Comput.-Aided Mol. Des.* **2007**, *21*, 421–435.
(54) Hochhaus, A.; La Rosee, P. Imatinib therapy in chronic myleogenous leukemia: Strategies to avoid and overcome resistance. *Leukemia* **2004**, *18*, 1321–31.
(55) (a) Nagar, B.; Bornmann, W. G.; Pellicena, P.; Schindler, T.; Veach, D. R.; Miller, W. T.; Clarkson, B.; Kuriyanet, 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. (b) Schindler, T.; Bornmann, W.; Pellicena, P.; Miller, T. W.; Clarkson, B.; Kuriyan, J. Structural mechanism of STI-571 inhibition of Abelson tyrosine kinase. *Science* **2000**, *289*, 1938–1942. (c) Nagar, B.; Bornmann, G. W.; Pellicena, P. 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.
(56) (a) Manley, W. P.; Cowan-Jacob, W. S.; Mestan, J. Advances in the structural biology, design and clinical development of Bcr-Abl kinase inhibitors for the treatment of chronic myeloid leukaemia. *Biochim. Biophys. Acta* **2005**, *1754*, 3–13. (b) Golas, M. J.; Arndt, K.; Etienne, C.; Lucas, J.; Nardin, D.; Boschelli1, D. H.; Boschelli1, F. SKI-606, A 4-anilino-3-quinolinecarbonitrile dual inhibitor of Src and Abl kinases, is a potent antiproliferative agent against chronic myelogenous leukemia cells in culture and causes regression of K562 xenografts in nude mice. *Cancer Res.* **2003**, *63*, 375–381. (c) Lombardo, L. J.; Lee, F. Y.; Chen, P.; Norris, D.; Barrish, J. C.; Behnia, K.; Castaneda, S.; Cornelius, L. A.; Das, J.; Doweyko, A. M.; Fairchild, C.; Hunt, J. T.; Inigo, I.; Johnston, K.; Kamath, A.; Kan, D.; Klei, H.; Marathe, P.; Pang, S.; Peterson, R.; Pitt, S.; Schieven, G. L.; Schmidt, R. J.; Tokarski, J.; Wen, M. L.; Wityak, J.; Borzilleri, R. M. Discovery of *N*-(2-chloro-6-methylphenyl)-2-(6-(4-(2-hydroxy-ethyl)-piperazin-1-yl)-2-methylpyrimidin-4-ylamino)thiazole-5-carboxamide (BMS-354825), A dual Src/Abl kinase inhibitor with potent antitumor activity in preclinical assays. *J. Med. Chem.* **2004**, *47*, 6658–6661.
(57) Strömbergsson, H.; Kryshtafovych, A.; Prusis, P.; Fidelis, K.; Wikberg, J. E. S.; Komorowski, J.; Hvidsten, T. R. Generalized modeling of enzyme−ligand interactions using proteochemometrics and local protein substructure. *Proteins* **2006**, *65*, 568–579.
(58) Vieth, M.; Higgs, R. E.; Robertson, D. H.; Shapiro, M.; Gragg, E. A.; Hemmerle, H. Kinomics-structural biology and chemogenomics of kinase inhibitors and targets. *Biochim. Biophys. Acta* **2004**, *1697*, 243–257.

CI1000532