Cross-Target View to Feature Selection: Identification of Molecular Interaction Features in Ligand−Target Space

Satoshi Niijima,* Hiroaki Yabuuchi, and Yasushi Okuno

Department of Systems Bioscience for Drug Discovery, Graduate School of Pharmaceutical Sciences, Kyoto University, Kyoto, Japan

Received April 12, 2010

There is growing interest in computational chemogenomics, which aims to identify all possible ligands of all target families using in silico prediction models. In particular, kernel methods provide a means of integrating compounds and proteins in a principled manner and enable the exploration of ligand−target binding on a genomic scale. To better understand the link between ligands and targets, it is of fundamental interest to identify molecular interaction features that contribute to prediction of ligand−target binding. To this end, we describe a feature selection approach based on kernel dimensionality reduction (KDR) that works in a ligand−target space defined by kernels. We further propose an efficient algorithm to overcome a computational bottleneck and thereby provide a useful general approach to feature selection for chemogenomics. Our experiment on cytochrome P450 (CYP) enzymes has shown that the algorithm is capable of identifying predictive features, as well as prioritizing features that are indicative of ligand preference for a given target family. We further illustrate its applicability on the mutation data of HIV protease by identifying influential mutated positions within protease variants. These results suggest that our approach has the potential to uncover the molecular basis for ligand selectivity and off-target effects.

INTRODUCTION

The last several years have seen a paradigm shift in pharmaceutical research from traditional target-specific approaches to a cross-target approach, offering tremendous opportunities for establishing novel drug design strategies to accelerate the drug discovery process. Receptors are no longer viewed as single entities but grouped into sets of related proteins or protein classes that are explored in a systematic manner. Chemogenomics has emerged as an interdisciplinary field, aiming at comprehensive coverage of ligand−target interactions, that is, identifying all possible ligands of all target families.[1,2] Concomitantly, high-throughput data being accumulated at an ever-increasing rate have triggered the development of novel in silico methodologies to comprehensively predict ligand−target interactions and binding affinities.[3]

In particular, a ligand−target approach has recently received much attention.[1] This approach represents a single-step process to integrate compounds and proteins into pairs and predict ligand−target binding on a genomic scale using machine learning models (e.g., that of Bock and Gough[4] and Erhan et al.[5]). The advantage of the ligand−target approach lies in that it allows predictions of new interactions even when neither ligands for a specific target nor targets for a specific ligand are known. Moreover, the greatest impact can be expected for targets devoid of structural 3D data, because classical drug design strategies like structure-based virtual screening cannot be applied to such targets.[1] Importantly, the ligand−target approach also has the potential to reveal ligand selectivity and off-target effects by comprehensive analysis of cross-reactivity of ligands.

Previous studies on the ligand−target approach have devoted much effort to the development of prediction models. Although advanced statistical models often yield better performance, they are usually constructed in a black-box way, lacking transparency and interpretability. This significantly hinders our understanding of the molecular basis for ligand−target binding. In order to gain an in-depth understanding of the link between ligands and targets, a next step should then be directed toward the identification of structural and physicochemical features associated with the binding. A promising in silico approach to this problem is to apply feature selection techniques,[6] which are typically used for molecular descriptor selection in chemoinformatics (e.g., the work of Fröhlich et al.,[7] Byvatov and Schneider,[8] and Xue et al.[9]). However, existing techniques for the ligand-based approach only consider individual targets and perform feature selection in the ligand space of a specific target and, thus, have a major limitation in capturing cross-reactive patterns. Given the fact that a single compound exhibits different binding affinities against multiple targets, feature selection needs to be performed instead in a ligand−target space, into which compounds are mapped jointly with targets. Because the ligand−target approach is itself an emerging strategy, feature selection based on the cross-target view is entirely an unexplored topic of research, and to our knowledge, no method exists that enables feature selection in the ligand−target space.

Here we describe a feature selection approach that works in a kernel-induced feature space[10] representing a ligand−target space. In particular, we propose using kernel dimensionality reduction (KDR)[11,12] for feature selection with an efficient

* To whom correspondence should be addressed. E-mail: niijima@pharm.kyoto-u.ac.jp.

10.1021/ci1001394
Published on Web 12/10/2010

algorithm, in order to identify molecular interaction features that contribute to prediction of ligand−target binding affinities. The quality of a prediction model is known to highly depend on the selected features and, hence, potentially benefits from feature selection. Indeed, the prediction performance can be improved by using only informative features. Reducing the number of features also helps to avoid overfitting.[13] Most importantly, selected features often facilitate interpretation of the model. For example, selected features in the ligand−target binding affinity space can serve to characterize privileged structures—selected substructures able to provide high-affinity ligands for a set of receptors[14]—and thus have implications for lead compound optimization for drug design. Furthermore, feature selection based on the cross-target view may provide insights into the molecular basis for ligand selectivity and off-target effects and has the potential to uncover the complex mode of drug actions. In the present study, we apply the proposed algorithm to a data set on cytochrome P450 (CYP) enzymes and show its capability of selecting a small subset of predictive features, which are further found to be indicative of ligand preference for a set of targets. We also evaluate our algorithm on the mutation data of HIV protease and illustrate its applicability by identifying influential amino acid positions within mutated variants.

METHODS

Representation of Ligand−Target Space. A key element of the ligand−target approach is the construction of ligand−target pairs, which need to be integrated from heterogeneous data types of compounds and proteins.[15] For this purpose, unified pair descriptions have been proposed and applied to search for novel active pairs.[4,5,16−21] In particular, it has recently proven that kernel methods[10] provide a general framework for integrating compounds and proteins, regardless of how they are represented, respectively.[17] Here we describe how a ligand−target space can be constructed via kernels.

Let us denote compounds and proteins by c_i and p_i, respectively. The binding affinity prediction problem can be formulated as the following machine learning problem: given a set of n ligand−target pairs (c_1,p_1), ..., (c_n,p_n) with known affinity values, construct a model to make predictions of activities of candidate pairs. To apply standard regression models, we first consider representing each ligand−target pair by a vector. Formally, given a chemical vector $\Phi_{\mathcal{C}}(c_i)$ and a protein vector $\Phi_t(p_i)$, we need to form a single vector $\Psi(c_i,p_i)$ using $\Phi_{\mathcal{C}}(c_i)$ and $\Phi_t(p_i)$.

To capture interactions between features of the compound c_i and those of the protein p_i, previous studies[5,16,17] proposed to represent the pair (c_i,p_i) by

$$\mathbf{\Psi}(c_i, p_i) = \mathbf{\Phi}_{\mathcal{C}}(c_i) \otimes \mathbf{\Phi}_t(p_i) \tag{1}$$

The tensor product operation $\otimes$ indicates that the pair is represented by all possible products (crossover) of the features of c_i and p_i, thereby seeking to fully encode correlations between them. The explicit computation of the products, however, demands expensive computation time and storage. Suppose that the number of features for c and p is d_c and d_p, respectively, then the pair is composed of $d_c \times d_p$ features. Fortunately, this computational bottleneck can be circumvented under the framework of kernel methods.[10]

Kernel methods are a class of algorithms that apply linear machine learning algorithms for classification or regression in a high-dimensional, possibly infinite-dimensional, feature space. Formally, the samples $x_i \in \mathrm{IR}^d$ are implicitly mapped into a feature space as $\Phi(x_i) \in \mathcal{F}$, such that the inner product between a pair of samples is given by a kernel function $k(x_i,x_j) = \Phi(x_i)\cdot\Phi(x_j)$, which measures the similarity between x_i and x_j. If the samples are expressed in terms of inner products only, the so-called kernel trick allows a variety of linear algorithms to work in the feature space constructed via a kernel function, without explicitly computing vectors comprising many features. In the case of binding affinity prediction, a ligand−target pair constitutes a single sample, and the kernel measures the similarity between ligand−target pairs. It can be shown that the kernel between pairs that are represented by eq 1 is decomposed as

$$\begin{aligned}\mathbf{\Psi}(c_i, p_i)^{\mathrm{T}}\mathbf{\Psi}(c_j, p_j) &= (\mathbf{\Phi}_{\mathcal{C}}(c_i) \otimes \mathbf{\Phi}_t(p_i))^{\mathrm{T}}(\mathbf{\Phi}_{\mathcal{C}}(c_j) \otimes \mathbf{\Phi}_t(p_j)) \\ &= \mathbf{\Phi}_{\mathcal{C}}(c_i)^{\mathrm{T}}\mathbf{\Phi}_{\mathcal{C}}(c_j) \times \mathbf{\Phi}_t(p_i)^{\mathrm{T}}\mathbf{\Phi}_t(p_j)\end{aligned} \tag{2}$$

It is readily seen that the similarity between two ligand−target pairs is simply the product of the similarity between the two compounds and the similarity between the two proteins. This indicates that eq 2, known as the tensor product kernel, can be computed easily once ligand and target kernels have been computed separately, avoiding the explicit computation of all the products of the features representing compounds and proteins. More generally, the kernels for compounds and proteins are not limited to the inner products of vectors and allow one to define the similarity based on nonvectorial data such as graphs for compounds, and amino acid sequences or 3D structures for proteins. Formally, denoting the kernels for compounds and proteins respectively by

$$k_{\mathcal{C}}(c_i, c_j) = \mathbf{\Phi}_{\mathcal{C}}(c_i)^{\mathrm{T}}\mathbf{\Phi}_{\mathcal{C}}(c_j) \tag{3}$$

$$k_t(p_i, p_j) = \mathbf{\Phi}_t(p_i)^{\mathrm{T}}\mathbf{\Phi}_t(p_j) \tag{4}$$

the kernel between the two pairs is given by

$$k((c_i, p_i), (c_j, p_j)) = k_{\mathcal{C}}(c_i, c_j) \times k_t(p_i, p_j)$$

In this way, the kernel-based approach allows versatile representation of the ligand−target space.

Feature Selection in Ligand−Target Binding Affinity Space. To select informative features in the ligand−target space as defined above, we need to perform feature selection in a kernel-induced feature space. Despite a broad spectrum of existing methods for feature selection,[6] there are few techniques that can be applied to such a ligand−target space constructed via kernels.

Here we adapt a semiparametric dimensionality reduction approach, called kernel dimensionality reduction (KDR),[11,12] to feature selection for binding affinity prediction. In particular, we propose an efficient feature selection algorithm to identify molecular interaction features that contribute to prediction of ligand−target binding.

KDR is a statistically grounded approach to dimensionality reduction, which aims to represent new features in the form of linear combinations of original features. This is achieved

by minimizing the independence (i.e., maximizing the dependence) between a set of features of samples and their labels. KDR can also be used to select a subset of original features that well captures the dependency. KDR enables us to measure the (in)dependence in a kernel-induced feature space, thereby providing a general means for dimensionality reduction and feature selection. In terms of statistics, KDR is based on the estimation and optimization of covariance operators on kernel-induced feature spaces, and the operators are used to provide a general characterization of conditional independence. KDR has the advantage that it imposes no strong assumptions either on the marginal distributions of samples and labels or on the conditional probability of labels given samples. This makes it applicable to diverse problems. Nevertheless, the application of KDR is still limited to typical machine learning problems[12] and yet to be seen in the chemoinformatics domain. Of note, this study is distinguished from others in that KDR is adapted to feature selection in a joint feature space of ligands and targets.

Among possible KDR objective functions, we employ the following simple function based on the trace of the empirical conditional covariance operator:[12]

$$\mathrm{Tr}[(\mathrm{HKH} + \lambda I_n)^{-1}\mathrm{HLH}] \tag{5}$$

Here, Tr denotes the trace of a matrix, and $\mathrm{K}, \mathrm{L} \in \mathrm{IR}^{n\times n}$ are the kernel matrices for the samples x_i and the labels y_i, respectively. $I_n \in \mathrm{IR}^{n\times n}$ is the identity matrix, and λ denotes a regularization parameter. $\mathrm{H} = I_n - (1/n)ee^{\mathrm{T}}$ is a centering matrix, where $e = (1, ..., 1)^{\mathrm{T}}$ is an n-dimensional vector. It is worth noting that this objective function has a close relationship with the Hilbert−Schmidt independence criterion (HSIC),[22] and eq 5 can be derived from the objective function of kernel ridge regression.[23] It is interesting to note that sliced inverse regression (SIR),[24] which is well-known and closely related to KDR, has recently been extended to kernel SIR (KSIR)[25] to overcome some limitations of SIR, yet unlike KDR, KSIR is sensitive to the number of slices which needs to be set a priori.

In the context of binding affinity prediction, the kernel matrix K defines the similarities between pairs, $x_i = (c_i,p_i)$, and L is simply computed as $L_{ij} = y_iy_j$, where y_i represents the affinity value given to x_i. Further, if we use the tensor product kernel eq 2, K can be represented as

$$K = K_{\mathscr{l}} \bigcirc K_t$$

where the elements of $K_{\mathscr{l}}$ and K_t are calculated by eqs 3 and 4 and $\bigcirc$ denotes the Hadamard product (elementwise product) operation.

As detailed in the work of Fukumizu et al.,[12] selection of relevant features exhibiting high dependence (i.e., low independence) on the labels reduces to the minimization of the objective function eq 5. Since exhaustive search is computationally prohibitive, we aim to achieve this with a backward elimination algorithm—the relevance of individual features is evaluated on a leave-one-out basis, and the least dependent feature maximizing the objective function is recursively eliminated from a full feature set. Alternative greedy algorithms such as forward search can also be used, but the backward elimination algorithm often yields better features, due to the evaluation of features in the presence of all others. To name but a few of this kind, SVM-RFE[26] and the BAHSIC algorithm[23] have indeed shown successful results.

Equation 5 can be computed independent of a particular classifier, yet the objective function involves the inverse of a sample-sized matrix. Thus, regardless of the search algorithm used, the computation of eq 5 based on leave-one-feature-out (LOFO) becomes intractable as the sample size and/or feature size increases.

Efficient Feature Selection Algorithm. To overcome this computational bottleneck, we propose an efficient algorithm for feature selection in the ligand−target binding affinity space constructed via the tensor product kernel. Specifically, we seek to improve the computational efficiency of

$$(\mathrm{H}(K_{\mathscr{l}} \bigcirc K_t)\mathrm{H} + \lambda I_n)^{-1}$$

in the LOFO process, i.e.,

$$\Delta^{(-i)} = (\mathrm{H}((K_{\mathscr{l}} - f_{\mathscr{l}}^{(i)}f_{\mathscr{l}}^{(i)\mathrm{T}}) \bigcirc K_t)\mathrm{H} + \lambda I_n)^{-1} \tag{6}$$

when selecting chemical features of ligands, while keeping protein features of targets unchanged. Here, $f_{\mathscr{l}}^{(i)}$ denotes a chemical feature to be left out. Note that the proposed algorithm is valid only when the linear kernel is used for ligands, but the targets can be represented by various features implicitly defined by kernels. First, we approximate the target kernel matrix K_t by a matrix G of lower-rank k as

$$K_t \approx \mathrm{GG}^{\mathrm{T}},\ \mathrm{G} = (g_1, ..., g_k) \in \mathrm{IR}^{n\times k} \tag{7}$$

This low-rank approximation can be efficiently done using, e.g., incomplete Cholesky decomposition.[27] For simplicity, let us define

$$\mathrm{P} = \mathrm{H}(K_{\mathscr{l}} \bigcirc K_t)\mathrm{H} + \lambda I_n \in \mathrm{IR}^{n\times n}$$

$$\mathrm{Q} = \mathrm{H}(f_{\mathscr{l}}^{(i)} \bigcirc g_1, ... f_{\mathscr{l}}^{(i)} \bigcirc g_k) \in \mathrm{IR}^{n\times k}$$

From eq 6, we have

$$\Delta^{(-i)} = (\mathrm{P} - \mathrm{QQ}^{\mathrm{T}})^{-1}$$

Further, from the Sherman−Morrison−Woodbury formula,[28] we have

$$\Delta^{(-i)} = \mathrm{P}^{-1} + \mathrm{P}^{-1}\mathrm{Q}(I_k - \mathrm{Q}^{\mathrm{T}}\mathrm{P}^{-1}\mathrm{Q})^{-1}\mathrm{Q}^{\mathrm{T}}\mathrm{P}^{-1}$$

Therefore, if $k \ll n$, computing the matrix inversion of $I_k - \mathrm{Q}^{\mathrm{T}}\mathrm{P}^{-1}\mathrm{Q} \in \mathrm{IR}^{k\times k}$ is efficient, and hence, $\Delta^{(-i)}$. Equation 7 can be computed independent of the LOFO process, and P^{-1} can be updated consecutively. In the case of binding affinity prediction, k is upper-bounded by $\min(n_p,d_p)$, where n_p is the number of proteins. Because binding affinities are typically measured for a relatively small number of targets against a series of compounds in chemical libraries, $k \le n_p \ll n$ usually holds, and the overall computation time can be saved by (n/k)-fold compared with a naive computation of eq 6. Of note, the proposed algorithm allows one to use different kernels for proteins when selecting chemical features. The algorithm can be summarized as follows:

Input: $\{((c_1,p_1),y_1),\ldots,((c_n,p_n),y_n)\}$,
chemical feature set $\mathscr{S}$

Output: A ranking list $\mathscr{R}$ of chemical features

1: Compute H, L, K_ℓ and K_t;

2: Compute G, Q, and P^{-1};

3: $\mathscr{R} \leftarrow \emptyset$;

Repeat 4–7 until $\mathscr{S} = \emptyset$

4: $j \leftarrow \arg\max_{i \in \mathscr{S}} \mathrm{Tr}\left[\Delta^{(-i)} HLH\right]$;

5: $P^{-1} \leftarrow \Delta^{(-j)}$;

6: $\mathscr{S} \leftarrow \mathscr{S} \setminus \{j\}$;

7: $\mathscr{R} \leftarrow \mathscr{R} \cup \{j\}$.

In the above algorithm, a single feature is recursively eliminated from $\mathscr{S}$ and added to the end of $\mathscr{R}$, in which the features toward the end of $\mathscr{R}$ have higher dependence on the labels in the presence of target information. Accordingly, the top-ranked features can be finally taken from the tail of $\mathscr{R}$.

Likewise, in the case of protein feature selection with chemical features unchanged, the same algorithm is applicable by simply replacing eqs 6 and 7 with

$$\Delta^{(-i)} = (\mathrm{H}((K_t - f_t^{(i)} f_t^{(i)\mathrm{T}}) \bigcirc K_\ell)\mathrm{H} + \lambda I_n)^{-1}$$

$$K_\ell \approx \mathrm{GG}^{\mathrm{T}}$$

where $f_t^{(i)}$ denotes a protein feature.

EXPERIMENTS

Data Sets. The CYP data set was taken from the study of Kontijevskis et al.[29] The affinity values of 798 ligand−target pairs (consisting of 371 inhibitors and 14 CYP enzymes) were experimentally determined and thus available. Each pair has a $\mathrm{pIC}_{50} = -\log(\mathrm{IC}_{50})$ value, where IC_{50} represent half-maximal inhibitory concentrations. The pIC_{50} values range from 0.46 to 8.70, with a mean value of 4.39. The distributions of the pIC_{50} values are shown for each CYP enzyme in Figure 1.

The mutation data of HIV protease were collected from the literature listed in the work of Lapins and Wikberg.[30] After carefully checking the literature sources, we chose to use a total of 389 ligand−target pairs with known $\mathrm{p}K_i$ values, where $\mathrm{p}K_i = -\log(K_i)$ and K_i represents inhibition constants. The ligand−target pairs consist of 21 ligands and 69 mutated protease variants as well as the wild-type, and the number of mutated positions amounts to 42 in the variants. The $\mathrm{p}K_i$ values range from 5.37 to 11.89, with a mean value of 8.75. The distributions of the $\mathrm{p}K_i$ values are shown for each ligand in Figure 2.

Kernels for Ligands and Targets. A wide variety of molecular features (descriptors) have been developed thus far to characterize the chemical structures and the physicochemical and molecular properties of compounds.[31] Here, we chose to use a total of 1664 descriptors calculated by version 1.4 of the Dragon software.[32] This descriptor set contains a range of 1D to 3D molecular features that fall into the following categories: constitutional descriptors, topological descriptors, walk and path counts, connectivity indices, information indices, 2D autocorrelations, edge adjacency indices, Burden eigenvalue descriptors, topological

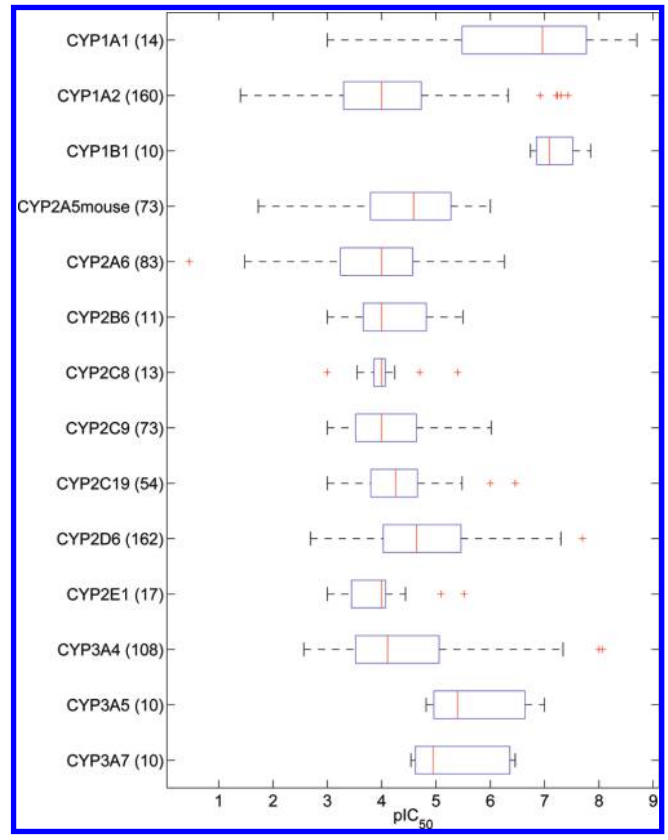

Figure 1. Boxplots of the pIC_{50} values for 14 CYP enzymes. Shown in parentheses are the numbers of inhibitors.

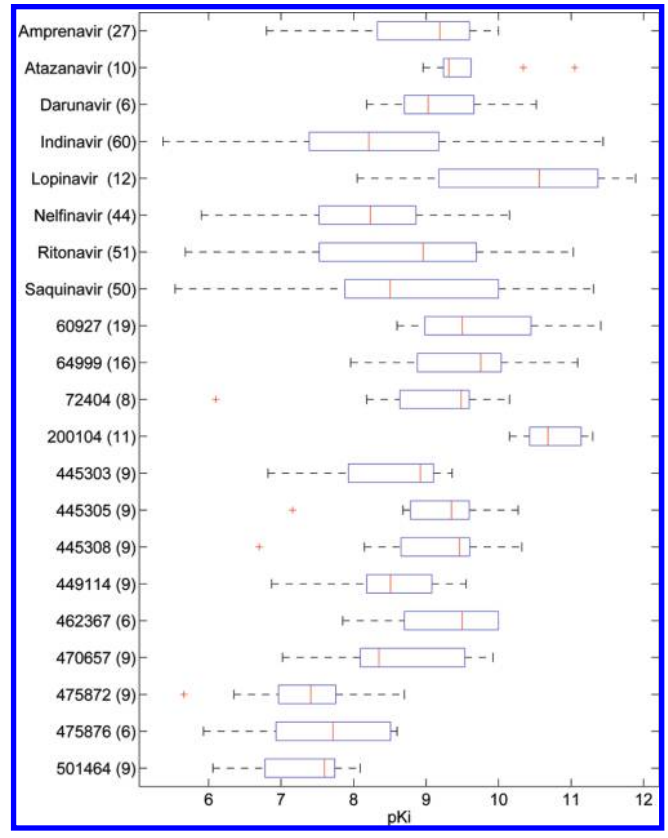

Figure 2. Boxplots of the $\mathrm{p}K_i$ values for 21 ligands. The ligand numbers indicate PubChem CIDs. Shown in parentheses are the numbers of mutated protease variants and the wild-type.

charge indices, eigenvalue-based indices, Randic molecular profiles, geometrical descriptors, RDF descriptors, 3D-MoRSE descriptors, WHIM descriptors, GETAWAY descriptors, functional group counts, atom-centered fragments, charge descriptors, and molecular properties. Before calculating these descriptors, MOE[33] was used to preprocess the raw macromolecular structures, including elimination of the crystallographic water molecules, removal of salts, addition of hydrogen atoms, and charge processing. A variation filter was then applied to eliminate the descriptors showing little variation across the compounds, resulting in 1397 descriptors for the CYP data set and 1378 descriptors for the HIV protease data set, and the values were scaled in the range of −1 to 1.

There exist several means of representing proteins or defining protein kernels. Among others, the sequence-based approach has proven effective when the availability of 3D structures is very limited. In view of this, we employed two different kernels in the experiment for CYPs: PROFEAT feature vectors[34] with RBF kernel (PROFEAT+RBF) and mismatch kernel.[35] These kernels can be computed from sequences alone and have shown good performance in protein classification and remote homology detection, as well as in ligand prediction.[17,20]

The PROFEAT feature vector provided by the PROFEAT Webserver[34] contains 1497 features representing, e.g., dipeptide composition and physicochemical properties of sequences. The RBF kernel was calculated using the feature vectors to represent the sequence similarity. The mismatch kernels are a class of string kernels, which can be computed as a dot product between two vectors consisting of frequencies of subsequences within the whole sequence. The mismatch kernels allow for mutations between the subsequences. Specifically, the mismatch kernel is calculated based on shared occurrences of (k,m)-patterns in the data, where the (k,m)-pattern consists of all k-length subsequences that differ from it by at most m mismatches. In our experiment, the typical choice of $k = 3$ and $m = 1$ was used in accordance with the work of Jacob and Vert.[17]

Whereas the chemical descriptors were subjected to feature selection for the CYP data set, feature selection was applied to protein features in the experiment for HIV protease, with the representation of ligands unchanged. Therefore, different kernels can be used for representing ligands, and we herein used the Dragon descriptors with RBF kernel. The targets were described using three z-scales, z_1, z_2, and z_3[36] following the work of Lapins and Wikberg.[30] The z-scales are the leading principal components obtained from 26 measured and computed physicochemical properties of amino acids and can be interpreted as hydrophobicity (z_1), steric properties (z_2), and polarity (z_3) of amino acids. As a result, the total number of protein features amounts to $42 \times 3 = 126$.

Performance Evaluation. The proposed algorithm selects features independent of a specific classifier used. It is therefore of interest to evaluate the predictive ability of the selected features using different kernel-based regression models. In the present study, we employed two representative models: kernel ridge regression (KRR)[37] and support vector regression (SVR).[38] The regularization parameter of KRR was fixed to the average eigenvalue of the kernel matrix. For SVR, the regularization parameter C was selected from {0.01, 0.1, ..., 100}, and the default value of 0.1 was used for ε of loss function. The γ parameter of RBF kernel for compounds (the CYP data set) and for proteins (the HIV protease data set) was set to α/(number of features), and α was selected from $\{2^{-4}, 2^{-3}, ..., 2^4\}$. The parameter λ of eq 5 was fixed to the average eigenvalue of HKH, which can easily be computed as Tr(HKH)/n. We used the LIBSVM library[39] for the implementation of SVR and in-house C codes for feature selection and KRR.

We used repeated random splitting for performance evaluation—the whole samples were partitioned randomly and repeatedly into training and test sets. The ratio of the training against test set was set to 6:1 for the CYP and HIV protease data sets, in accordance with previous studies.[29,30] Feature selection was performed using only the training set, and the q^2 value of the learnt regression model was obtained using the test set. Given n test samples, q^2 is defined as

$$q^2 = 1 - \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

where y_i and $\hat{y}_i$ are the true affinity value and estimated value of sample i, respectively, and $\bar{y}$ is the average value of y_is. Thus, the larger value of q^2 indicates the better performance. The random splitting was repeated 20 times, and the q^2 value averaged over the 20 runs and the corresponding standard deviations are reported here. Because the rank of features can vary depending on the training sets, we calculated scores as the average ranks of the 20 ranking lists.

The aim of the experiments was to evaluate how the prediction performance would be affected by eliminating possibly irrelevant features and whether our algorithm can identify a small set of informative features. To this end, the number of features was varied from all features to >50 by 10% decrements, and from 50 to 5 in decrements of 5. The predictive ability of the selected features was assessed by KRR and SVR as a function of the number of features. There exists no competing method that enables feature selection in a ligand−target space constructed via kernels, but it is worth making a comparison with random selection as a baseline to evaluate how well the proposed algorithm performs in practice. For this purpose, we randomly selected features from the whole feature set for each data set splitting and evaluated their prediction performance in the same way as for the selected features of the proposed algorithm.

RESULTS AND DISCUSSION

Chemical Feature Selection for CYPs. CYPs constitute a superfamily of heme-containing enzymes, which are involved in the oxidative metabolism of a large number of structurally different compounds of both endogenous and exogenous origin. It is known that more than 90% of all pharmaceuticals are metabolized by CYPs; CYP1A2, CYP2C9, CYP2C19, CYP2D6, and CYP3A4 are predominant among others.[40] These enzymes are susceptible to inhibition due to their broad specificity, giving rise to unexpected drug−drug interactions and drug toxicity. This makes prediction of interactions between CYPs and drugs a challenging problem.

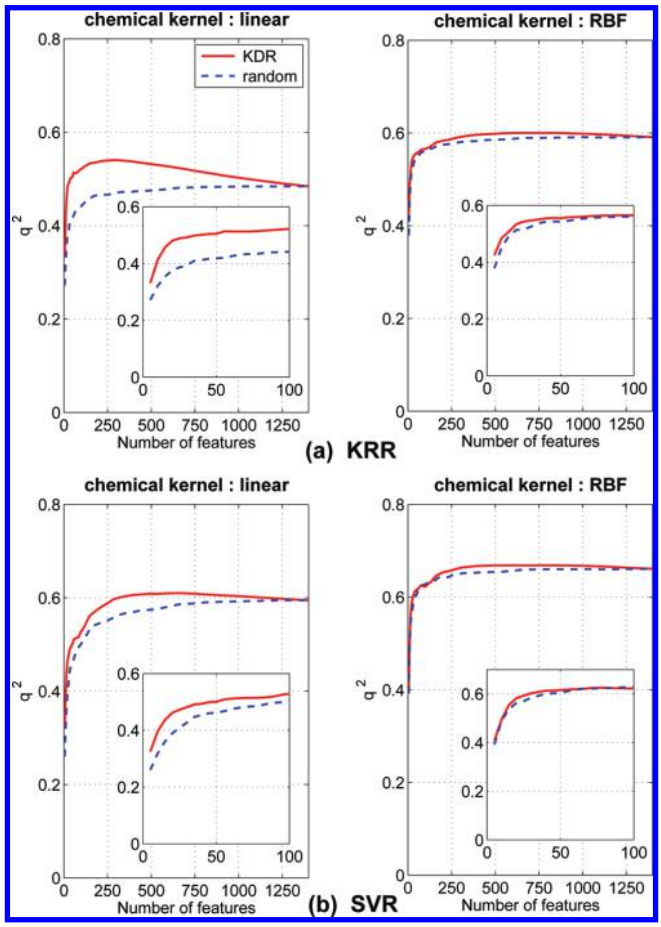

Figure 3. Average q^2 values as a function of the number of features. The PROFEAT+RBF kernel was used for CYPs. (a) KRR with the linear and RBF kernels for compounds. (b) SVR with the linear and RBF kernels for compounds.

As shown in Figure 1, the distributions of the pIC_{50} values significantly overlap between the predominant CYPs and exhibit a wide range of inhibitory activities for most of the CYP enzymes, albeit biased for a fraction of them (e.g., CYP1B1). When simply estimating the value of a given inhibitor to be the mean value for the target CYP, we observed $q^2 = 0.18 \pm 0.07$, which clearly indicates the need for the use of both ligand and target information to make better predictions, and this can be achieved by the ligand−target approach.

We applied the proposed algorithm to the CYP data set. Figure 3 and Table 1 show the q^2 values for KRR and SVR with the linear and RBF kernels for compounds and the PROFEAT+RBF kernel for CYPs. Using all the chemical features, KRR and SVR with the RBF kernel yielded $q^2 =$ 0.59 and 0.66, respectively. The same data set was analyzed by Kontijevskis et al.[29] using a different ligand−target approach that is based on linear partial least-squares (PLS), and the PLS-based model yielded q^2 values of 0.61−0.66. Although a fair comparison of the performance with the present study cannot be made due to the difference in chemical descriptors used, $q^2 = 0.66$ obtained by our SVR is comparable to the reported values in the previous study. It should be noted that the q^2 value exceeding 0.60 can be considered highly predictive, compared with previous in silico models for predicting CYP inhibition.[29]

To evaluate whether our algorithm can narrow an abundance of features that possibly include irrelevant ones down to the most informative features, we compared the performance between the proposed algorithm based on KDR and random selection by varying the number of chemical features. As the chemical features were removed, the q^2 values for random selection dropped gradually, whereas our algorithm was able to reduce the number of features while maintaining the same level of performance. As seen for KRR with the linear kernel, the performance could even be improved by reducing the feature size (Table 1); however, this was not observed for random selection. This result suggests that only a small subset of features is sufficient for making accurate prediction, while most of the other features are likely irrelevant to the prediction. Indeed, it can be seen from Table 1 that the q^2 value for SVR with the RBF kernel decreased merely from 0.66 with all features to 0.62 with 50 features.

Table 1. Performance Comparison for CYPs Using the PROFEAT+RBF kernel[a]

	KRR (chemical kernel: linear)		KRR (chemical kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.41 ± 0.08	0.32 ± 0.09	0.48 ± 0.07	0.45 ± 0.08
20	0.48 ± 0.07	0.38 ± 0.09	0.53 ± 0.06	0.51 ± 0.08
30	0.49 ± 0.07	0.39 ± 0.09	0.55 ± 0.07	0.53 ± 0.08
50	0.50 ± 0.07	0.42 ± 0.10	0.55 ± 0.07	0.54 ± 0.07
108	0.52 ± 0.08	0.44 ± 0.10	0.57 ± 0.06	0.56 ± 0.07
all (1397)	0.48 ± 0.08		0.59 ± 0.07	
	0.54 ± 0.07 (KDR, 185)		**0.60** ± 0.06 (KDR, 392)	
	SVR (chemical kernel: linear)		SVR (chemical kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.40 ± 0.09	0.32 ± 0.12	0.50 ± 0.10	0.49 ± 0.11
20	0.46 ± 0.08	0.39 ± 0.12	0.58 ± 0.08	0.56 ± 0.11
30	0.48 ± 0.10	0.43 ± 0.12	0.60 ± 0.08	0.58 ± 0.10
50	0.50 ± 0.09	0.46 ± 0.12	0.62 ± 0.07	0.60 ± 0.09
108	0.53 ± 0.09	0.51 ± 0.12	0.63 ± 0.07	0.63 ± 0.09
all (1397)	0.59 ± 0.09		0.66 ± 0.09	
	0.61 ± 0.09 (KDR, 392)		**0.67** ± 0.07 (KDR, 352)	

[a] Shown are the average q^2 values and standard deviations. The last row shows the best values in bold face (with method and the number of features in parentheses).

Overall, the proposed algorithm performs better than random selection, but the difference is less remarkable for the RBF kernel. This may be because the features were optimized in the ligand−target space with the linear kernel for compounds and, hence, are not necessarily optimal in the ligand−target space with the RBF kernel. In principle, KDR can be computed in the latter space as well, but the computational cost is so demanding that our efficient algorithm can be a compromise between the cost and performance.

Because our feature selection algorithm is amenable to various kernels for proteins when selecting chemical features, the mismatch kernel was also applied in the same way. As shown in Figure 4 and Table 2, a similar trend was observed as the features were removed. In particular, we confirmed again that a small subset of features was as predictive as the given full feature set.

In the context of binding affinity prediction, the minimization of KDR favors chemical features that exhibit high dependence on affinity values in the presence of protein information. It is therefore of interest to see whether the

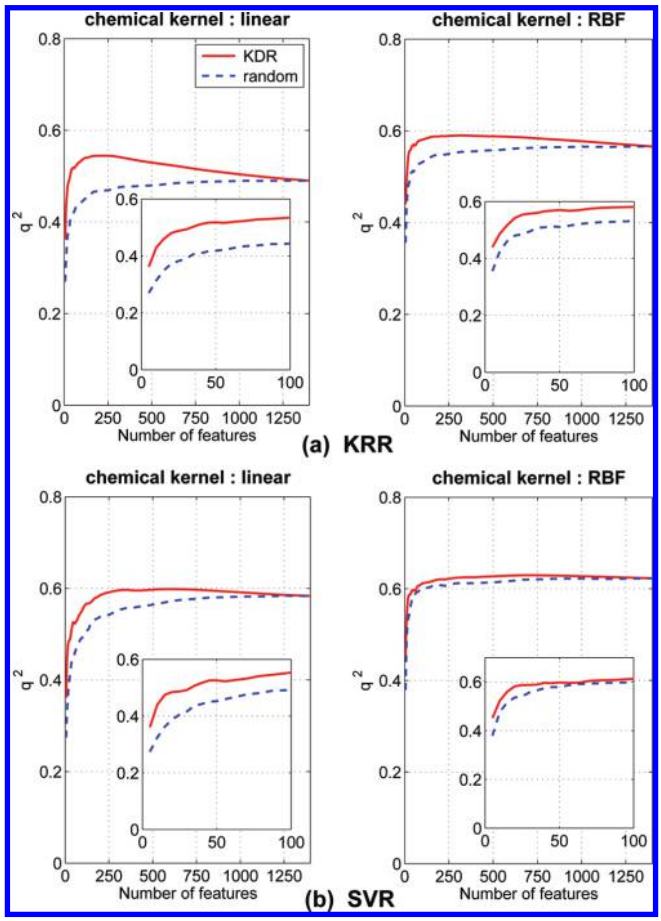

Figure 4. Average q^2 values as a function of the number of features. The mismatch kernel was used for CYPs. (a) KRR with the linear and RBF kernels for compounds. (b) SVR with the linear and RBF kernels for compounds.

Table 2. Performance Comparison for CYPs Using the Mismatch Kernel[a]

	KRR (chemical kernel: linear)		KRR (chemical kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.43 ± 0.08	0.31 ± 0.09	0.49 ± 0.08	0.43 ± 0.08
20	0.48 ± 0.06	0.37 ± 0.08	0.54 ± 0.07	0.48 ± 0.08
30	0.49 ± 0.07	0.39 ± 0.08	0.56 ± 0.07	0.49 ± 0.07
50	0.52 ± 0.08	0.42 ± 0.08	0.57 ± 0.07	0.51 ± 0.08
108	0.54 ± 0.07	0.45 ± 0.08	0.58 ± 0.07	0.53 ± 0.07
all (1397)	0.49 ± 0.07		0.57 ± 0.07	
	0.54 ± 0.07 (KDR, 108)		**0.59** ± 0.07 (KDR, 134)	
	SVR (chemical kernel: linear)		SVR (chemical kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.44 ± 0.08	0.32 ± 0.11	0.52 ± 0.10	0.48 ± 0.09
20	0.48 ± 0.07	0.39 ± 0.10	0.58 ± 0.08	0.54 ± 0.10
30	0.49 ± 0.08	0.42 ± 0.10	0.59 ± 0.08	0.56 ± 0.09
50	0.53 ± 0.09	0.45 ± 0.10	0.60 ± 0.08	0.58 ± 0.08
108	0.56 ± 0.09	0.50 ± 0.10	0.61 ± 0.08	0.60 ± 0.08
all (1397)	0.58 ± 0.08		0.62 ± 0.08	
	0.60 ± 0.09 (KDR, 316)		**0.63** ± 0.08 (KDR, 352)	

[a] Shown are the average q^2 values and standard deviations. The last row shows the best values in bold face (with method and the number of features in parentheses).

selected features can give some explanation about the characteristics of CYP inhibitors. Table 3 lists the top 30 features selected from a total of 1397 chemical features. It can be seen that two features representing the octanol−water partition coefficient (logP) received relatively high ranks (6th and 21st). Given that the ligand−target pairs for CYP1A2 and CYP3A4 account for more than 30% of the data set, this is in line with the fact that inhibitors of CYP1A2 and CYP3A4 are known to show high lipophilicity.[41,42] ARR (10th) and nBnz (23rd) are likely to reflect that CYP1A2 inhibitors have high aromaticity (number of aromatic carbons).[41] Also, aromatic groups such as pyridines, imidazoles, and phenols have been reported to characterize CYP3A4 inhibitors.[43,44] In addition, charge descriptors, qnmax (9th), qpmax (14th), RPCG (16th), and RNCG (17th), are indicative of the involvement of polarizability in CYP3A4 inhibitors.[42] These observations suggest that increasing lipophilicity, aromaticity, and polarizability would enhance inhibitory activity.

Taken together, our approach is capable of identifying predictive features, as well as prioritizing features that are characteristic of CYP inhibition. Since the feature set used contains many features that are not easily interpretable, it is difficult to fully explain the relevance of the selected features. Nevertheless, predictive features may serve as markers for triaging compounds with desired affinities. In light of interpretability, more elaborate description of structural features of compounds, such as extended connectivity fingerprints[45] may be preferred to the Dragon descriptors. To explore the predictive ability of such fingerprints, we also tested ECFP6 and ECFC6 (as calculated by Pipeline Pilot[46]) for the CYP data set. However, we found that ECFP and ECFC were less predictive than the Dragon descriptors and that physicochemical and molecular properties of compounds are better suited to predict the binding affinities of CYP inhibitors.

Protein Feature Selection for Mutated HIV Protease Variants. The proposed algorithm was also evaluated on the mutation data of HIV protease, a major target for highly active antiretroviral therapy. The ability of the HIV virus to mutate and develop drug resistance by accumulating mutations severely hinders the treatment of HIV. To guide the design of new inhibitors that surmount the resistance, it is of great value to understand the mutational determinants involved in the interactions between inhibitors and HIV protease variants. The composite effects of distantly located mutations and the phenomenon of cross-resistance further motivate us to explore the mutational space of the protease in a comprehensive manner.[47]

As shown in Figure 2, the distributions of the pK_i values are wide-ranging to varying degrees and heavily overlap among one another. This suggests that simply estimating the value of a given protease variant to be the mean value for the ligand of interest is unsatisfactory ($q^2 = 0.21 \pm 0.09$) and that both target and ligand information is needed for accurate predictions.

Figure 5 and Table 4 show the q^2 values for KRR and SVR with the RBF kernel for compounds and the linear and RBF kernels for mutated HIV protease variants. Using all the protein features, KRR and SVR with the RBF kernel yielded $q^2 = 0.70$ and 0.78, respectively. A q^2 value of 0.78 is quite consistent with the best q^2 values of 0.78−0.83 reported in the study of Lapins andWikberg,[30] despite some differences in the data set and descriptors used.

Table 3. Top-Ranked Chemical Features of CYP Inhibitors[a]

rank	symbol	description	score
1	piPC09	molecular multiple path count of order 09	1.70
2	MATS1v	Moran autocorrelation−lag 1/weighted by atomic van der Waals volumes	5.15
3	Hypertens-80	Ghose−Viswanadhan−Wendoloski antihypertensive-like index at 80%	5.90
4	BIC0	bond information content (neighborhood symmetry of 0-order)	6.05
5	Infective-80	Ghose−Viswanadhan−Wendoloski antiinfective-like index at 80%	8.45
6	ALOGP2	Squared Ghose−Crippen octanol−water partition coeff ($logP^2$)	9.20
7	RARS	R matrix average row sum	9.60
8	G3s	third component symmetry directional WHIM index/weighted by atomic electrotopological states	13.25
9	qnmax	maximum negative charge	13.40
10	ARR	aromatic ratio	14.60
11	GATS3v	Geary autocorrelation−lag 3/weighted by atomic van der Waals volumes	15.35
12	MATS1p	Moran autocorrelation−lag 1/weighted by atomic polarizabilities	16.40
13	BEHp6	highest eigenvalue *n*. 6 of Burden matrix/weighted by atomic polarizabilities	16.80
14	qpmax	maximum positive charge	19.00
15	C-015	$=CH_2$	20.00
16	RPCG	relative positive charge	20.70
17	RNCG	relative negative charge	21.15
18	REIG	first eigenvalue of the R matrix	21.50
19	GATS1m	Geary autocorrelation−lag 1/weighted by atomic masses	22.35
20	R3e+	R maximal autocorrelation of lag 3/weighted by atomic Sanderson electronegativities	24.95
21	ALOGP	Ghose−Crippen octanol−water partition coeff (logP)	27.00
22	BLTF96	Verhaar model of Fish baseline toxicity for Fish (96 h) from MLOGP (mmol/L)	27.15
23	nBnz	number of benzene-like rings	28.55
24	Mor23v	3D-MoRSE−signal 23/weighted by atomic van der Waals volumes	28.95
25	SIC0	structural information content (neighborhood symmetry of 0-order)	30.05
26	BEHm7	highest eigenvalue *n*. 7 of Burden matrix/weighted by atomic masses	31.80
27	BEHv6	highest eigenvalue *n*. 6 of Burden matrix/weighted by atomic van der Waals volumes	32.70
28	GATS3p	Geary autocorrelation−lag 3/weighted by atomic polarizabilities	32.85
29	BLTA96	Verhaar model of Algae baseline toxicity for Algae (96 h) from MLOGP (mmol/L)	33.60
30	G3e	third component symmetry directional WHIM index/weighted by atomic Sanderson electronegativities	34.25

[a] The PROFEAT+RBF kernel was used for CYPs.

We then compared the performance between the proposed algorithm based on KDR and random selection with varying numbers of protein features. As the protein features were removed, the q^2 values for random selection dropped markedly. In contrast, the proposed algorithm successfully reduced the number of features to less than 30 while maintaining high q^2 values, clearly outperforming random selection. Indeed, in the case of SVR with the RBF kernel, the performance slightly drops from $q^2 = 0.78$ to 0.73 using the top 30 features of KDR, yet this value is significantly higher than $q^2 = 0.60$ obtained by random selection.

The selected protein features are those exhibiting high dependence on affinity values in the presence of chemical information, and the performance curve in Figure 5 suggests the biological relevance of the top 20−30 features. Table 5 lists the 20 top-ranked mutated positions with amino acid properties (z-scales). This list indicates that the most influential positions are 36, 48, 50, 63, 82, 84, and 90. Indeed, positions 48, 50, 82, and 84 are known as the active site of the protease.[47] It is thus likely that mutating these positions has a great effect on decreasing inhibitory activity of a group of inhibitors. The proposed algorithm is a multivariate approach and hence can detect composite effects of multi-mutations. Consistent with this, positions 48, 82, 84, and 90 have been identified as being interrelated with each other.[30] Interestingly, position 90 is located in the dimerization region of the protease, but such a distantly located mutation has been shown to prevent ligands from binding the protease by changing the geometry of the active site.[48] On the other hand, positions 36 or 63 are prone to natural genetic variations and may not by themselves confer resistance to inhibitors.[49] Our analysis identified them as informative, and this may be due to composite effects with other mutated positions, an observation that has also been suggested in the previous study.[30] For most of the top-ranked mutated positions, all the three amino acid properties seem to be relevant, but this is not the case for position 82. Specifically, 82 (z_1) and 82 (z_3) representing hydrophobicity and polarity were ranked the 14th and 9th, respectively, whereas 82 (z_2) representing steric properties was ranked the 48th and hence less relevant. This is also in good agreement with the previous study,[30] but the relevance of other top-ranked mutations such as 37 (z_2) and 71 (z_3) remains elusive. Overall, these results illustrate that the proposed feature selection approach serves as a useful tool not only for identifying informative chemical

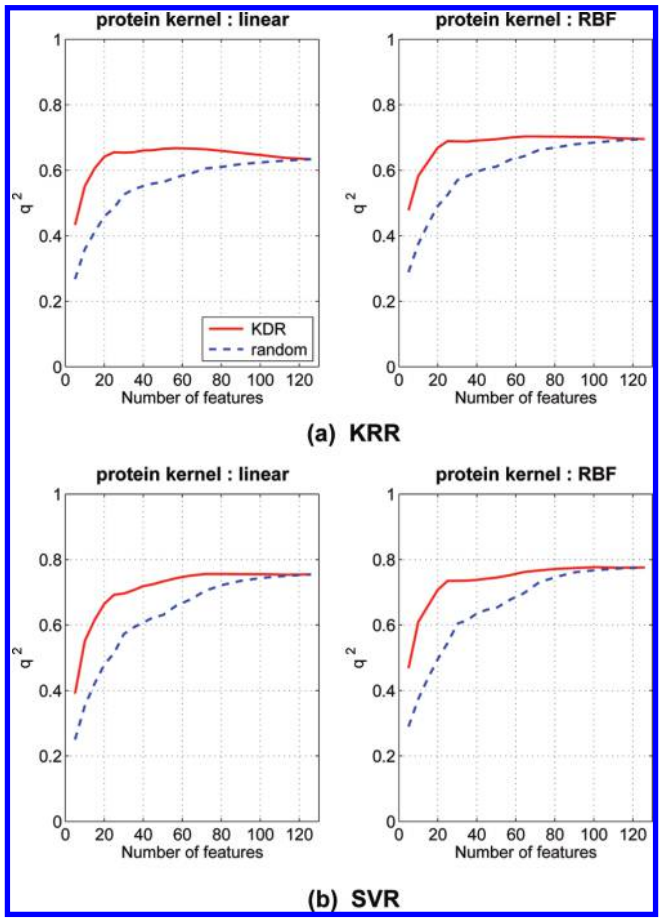

Figure 5. Average q^2 values as a function of the number of features. The RBF kernel was used for mutated HIV protease variants. (a) KRR with the linear and RBF kernels for compounds. (b) SVR with the linear and RBF kernels for compounds.

Table 4. Performance Comparison for Mutated HIV Protease Variants[a]

	KRR (protein kernel: linear)		KRR (protein kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.55 ± 0.08	0.36 ± 0.12	0.58 ± 0.08	0.38 ± 0.13
20	0.64 ± 0.07	0.46 ± 0.12	0.67 ± 0.07	0.49 ± 0.12
30	0.65 ± 0.06	0.53 ± 0.09	0.69 ± 0.06	0.57 ± 0.10
all (126)	0.63 ± 0.06		**0.70** ± 0.05	
	0.67 ± 0.06 (KDR, 50)		**0.70** ± 0.06 (KDR, 50)	
	SVR (protein kernel: linear)		SVR (protein kernel: RBF)	
no. of features	KDR	random	KDR	random
10	0.55 ± 0.11	0.35 ± 0.14	0.61 ± 0.08	0.37 ± 0.15
20	0.66 ± 0.08	0.48 ± 0.13	0.71 ± 0.07	0.50 ± 0.13
30	0.70 ± 0.07	0.57 ± 0.12	0.73 ± 0.06	0.60 ± 0.12
all (126)	0.75 ± 0.04		**0.78** ± 0.04	
	0.76 ± 0.05 (KDR, 72)		**0.78** ± 0.04 (KDR, 101)	

[a] Shown are the average q^2 values and standard deviations. The last row shows the best values in bold face (with method and the number of features in parentheses).

features but also for analyzing the effect of multimutations on their affinity to a series of inhibitors. Importantly, the selected positions and properties have implications for engineering new mutations at the same positions.

Table 5. Top-Ranked Protein Features of Mutated HIV Protease Variants

rank	mutated position (*z*-scale)	score	rank	mutated position (*z*-scale)	score
1	90 (z_3)	1.00	11	71 (z_3)	10.15
2	36 (z_3)	2.00	12	50 (z_2)	11.95
3	84 (z_1)	4.60	13	54 (z_3)	12.40
4	63 (z_3)	4.85	14	82 (z_1)	13.20
5	84 (z_3)	6.90	15	90 (z_2)	15.20
6	50 (z_3)	8.20	16	50 (z_1)	15.50
7	36 (z_2)	8.35	17	30 (z_3)	18.15
8	48 (z_3)	8.55	18	84 (z_2)	19.20
9	82 (z_3)	8.95	19	37 (z_3)	19.35
10	37 (z_2)	9.75	20	46 (z_2)	19.75

CONCLUSION

We have proposed an efficient feature selection algorithm based on KDR to identify molecular interaction features that contribute to prediction of ligand−target binding. Unlike existing feature selection techniques for chemoinformatics, our approach performs chemical (protein) feature selection coupled with protein (compound) information. In particular, the proposed algorithm works in a ligand−target space defined by kernels, allowing one to use various kernels for proteins (compounds) in selecting chemical (protein) features and, thus, provides a useful general approach to feature selection for chemogenomics.

The experiment on CYPs has shown that the algorithm is capable of identifying predictive features, as well as prioritizing features that are indicative of ligand preference for a given target family. Notably, using only the relevant features can lead to an improved performance. We have further illustrated the applicability on the mutation data of HIV protease by identifying influential amino acid positions within mutated variants. These results suggest that our feature selection approach based on the cross-target view can not only aid in drug design but also provide clues as to the molecular basis for ligand selectivity and off-target effects. We envision that this study will encourage further research in computational chemogenomics and contribute to a better understanding of the mechanism of molecular recognition.

ACKNOWLEDGMENT

This work was supported by grants from the Ministry of Education, Culture, Sports, Science and Technology, Japan; the Targeted Proteins Research Program; and the New Energy and Industrial Technology Development Organization (NEDO) under the Ministry of Economy Trade and Industry of Japan. Financial support from Ono Pharmaceutical Co., Ltd., is also gratefully acknowledged.

REFERENCES AND NOTES

(1) Klabunde, T. Chemogenomic approaches to drug discovery: similar receptors bind similar ligands. *Br. J. Pharmacol.* **2007**, *152*, 5–7.
(2) Rognan, D. Chemogenomic approaches to rational drug design. *Br. J. Pharmacol.* **2007**, *152*, 38–52.
(3) Bajorath, J. Computational analysis of ligand relationships within target families. *Curr. Opin. Chem. Biol.* **2008**, *12*, 352–358.
(4) Bock, J. R.; Gough, D. A. Virtual screen for ligands of orphan G protein-coupled receptors. *J. Chem. Inf. Model* **2005**, *45*, 1402–1414.
(5) Erhan, D.; L'Heureux, P.-J.; Yue, S. Y.; Bengio, Y. Collaborative filtering on a family of biological targets. *J. Chem. Inf. Model.* **2006**, *46*, 626–635.

(6) Guyon, I.; Elisseeff, A. An introduction to variable and feature selection. *J. Mach. Learn. Res.* **2003**, *3*, 1157–1182.
(7) Fröhlich, H.; Wegner, J. K.; Zell, A. Towards optimal descriptor subset selection with support vector machines in classification and regression. *QSAR Comb. Sci.* **2004**, *23*, 311–318.
(8) Byvatov, E.; Schneider, G. SVM-based feature selection for characterization of focused com- pound collections. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 993–999.
(9) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of molecular de-scriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1630–1638.
(10) Schölkopf, B.; Smola, A. J. *Learning with Kernels: Support Vector Machines, Regulariza-tion, Optimization, and Beyond*; MIT Press: Cambridge, MA, 2002.
(11) Fukumizu, K.; Bach, F. R.; Jordan, M. I. Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces. *J. Mach. Learn. Res.* **2004**, *5*, 73–99.
(12) Fukumizu, K.; Bach, F. R.; Jordan, M. I. Kernel dimensionality reduction in regression. *Ann. Stat.* **2009**, *37*, 1871–1905.
(13) Hawkins, D. M. The problem of overfitting. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1–12.
(14) Schnur, D. M.; Hermsmeier, M. A.; Tebben, A. J. Are target-family-privileged substructures truly privileged. *J. Med. Chem.* **2006**, *49*, 2000–2009.
(15) Schuffenhauer, A.; Jacoby, E. Annotating and mining the ligand−target chemogenomics knowledge space. *Drug Discovery Today* **2004**, *2*, 190–200.
(16) Faulon, J.-L.; Misra, M.; Martin, S.; Sale, K.; Sapra, R. Genome scale enzyme-metabolite and drug-target interaction predictions using the signature molecular descriptor. *Bioinformatics* **2008**, *24*, 225–233.
(17) Jacob, L.; Vert, J.-P. Protein-ligand interaction prediction: an improved chemogenomics ap-proach. *Bioinformatics* **2008**, *24*, 2149–2156.
(18) Nagamine, N.; Sakakibara, Y. Statistical prediction of protein-chemical interactions based on chemical structure and mass spectrometry data. *Bioinformatics* **2007**, *23*, 2004–2012.
(19) Strömbergsson, H.; Daniluk, P.; Kryshtafovych, A.; Fidelis, K.; Wikberg, J. E.; Kley-wegt, G. J.; Hvidsten, T. R. Interaction model based on local protein substructures generalizes to the entire structural enzyme-ligand space. *J. Chem. Inf. Model* **2008**, *48*, 2278–88.
(20) Wassermann, A. M.; Geppert, H.; Bajorath, J. Ligand prediction for orphan targets using support vector machines and various target-ligand kernels is dominated by nearest neighbor effects. *J. Chem. Inf. Model* **2009**, *49*, 2155–2167.
(21) Weill, N.; Rognan, D. Development and validation of a novel protein-ligand fingerprint to mine chemogenomic space: application to G protein-coupled receptors and their ligands. *J. Chem. Inf. Model* **2009**, *49*, 1049–62.
(22) Gretton, A.; Bousquet, O.; Smola, A. J.; Schölkopf, B. *Proceedings of the Sixteenth International Conference on Algorithmic Learning Theory*; Singapore, Oct 8−11; Springer: Berlin/Heidelberg, 2005; pp 63−78.
(23) Song, L.; Bedo, J.; Borgwardt, K. M.; Gretton, A.; Smola, A. Gene selection via the BAHSIC family of algorithms. *Bioinformatics* **2007**, *23*, i490–i498.
(24) Li, K.-C. Sliced inverse regression for dimension reduction. *J. Am. Stat. Assoc.* **1991**, *86*, 316–327.
(25) Yeh, Y.-R.; Huang, S.-Y.; Lee, Y.-J. Nonlinear dimension reduction with kernel sliced inverse regression. *IEEE Trans. Knowledge Data Eng.* **2009**, *21*, 1590–1603.
(26) Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for cancer classification using support vector machines. *Mach. Learn.* **2002**, *46*, 389–422.
(27) Bach, F. R.; Jordan, M. I. Kernel independent component analysis. *J. Mach. Learn. Res.* **2002**, *3*, 1–48.
(28) Golub, G. H.; Loan, C. F. V. *Matrix Computations*, 3rd ed.; Johns Hopkins University Press: Baltimore, 1996.
(29) Kontijevskis, A.; Komorowski, J.; Wikberg, J. E. S. Generalized proteochemometric model of multiple cytochrome P450 enzymes and their inhibitors. *J. Chem. Inf. Model* **2008**, *48*, 1840–1850.
(30) Lapins, M.; Wikberg, J. E. S. Proteochemometric modeling of drug resistance over the mu-tational space for multiple HIV protease variants and multiple protease inhibitors. *J. Chem. Inf. Model* **2009**, *49*, 1202–1210.
(31) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.
(32) *DragonX*, version 1.2; Milano Chemometrics and QSAR Research Group: Milan, 2007.
(33) *MOE*, version 2008.10; Chemical Computing Group Inc.: Montreal, Canada, 2008.
(34) Li, Z. R.; Lin, H. H.; Han, L. Y.; Jiang, L.; Chen, X.; Chen, Y. Z. PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence. *Nucleic Acids Res.* **2006**, *34*, W32–W37.
(35) Leslie, C. S.; Eskin, E.; Cohen, A.; Weston, J.; Noble, W. S. Mismatch string kernels for discriminative protein classification. *Bioinformatics* **2004**, *20*, 467–476.
(36) Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J. Med. Chem.* **1998**, *41*, 2481–2491.
(37) Saunders, C.; Gammerman, A.; Vovk, V. *Proceedings of the Fifteenth International Conference on Machine Learning*; Wisconsin, July 24−27; Morgan Kaufmann Publishers, Inc.: San Francisco, 1998; pp 515−521.
(38) Vapnik, V. N. *Statistical Learning Theory*; John Wiley & Sons, Inc.: New York, 1998.
(39) Chang, C.-C.; Lin, C.-J. LIBSVM: a library for support vector machines. http://www.csie.ntu.edu.tw/∼cjlin/libsvm/.
(40) de Groot, M. J. Designing better drugs: predicting cytochrome P450 metabolism. *Drug Discovery Today* **2006**, *11*, 601–606.
(41) Chohan, K. K.; Paine, S. W.; Mistry, J.; Barton, P.; Davis, A. M. A rapid computational filter for cytochrome P450 1A2 inhibition potential of compound libraries. *J. Med. Chem.* **2005**, *48*, 5154–5161.
(42) Kriegl, J. M.; Arnhold, T.; Beck, B.; Fox, T. Prediction of human cytochrome P450 inhibition using support vector machines. *QSAR Comb. Sci.* **2005**, *24*, 491–502.
(43) Jensen, B. F.; Vind, C.; Padkjær, S. B.; Brockhoff, P. B.; Refsgaard, H. H. F. In silico predic-tion of cytochrome P450 2D6 and 3A4 inhibition using Gaussian kernel weighted k-nearest neighbor and extended connectivity fingerprints, including structural fragment analysis of inhibitors versus noninhibitors. *J. Med. Chem.* **2007**, *50*, 501–511.
(44) Riley, R. J.; Parker, A. J.; Trigg, S.; Manners, C. N. Development of a generalized, quanti-tative physicochemical model of CYP3A4 inhibition for use in early drug discovery. *Pharm. Res.* **2001**, *18*, 652–655.
(45) Rogers, D.; Hahn, M. Extended-connectivity fingerprints. *J. Chem. Inf. Model* **2010**, *50*, 742–754.
(46) *Pipeline Pilot*, version 6.5.1; Accelrys, Inc.: San Diego, CA, 2007.
(47) Lapins, M.; Eklund, M.; Spjuth, O.; Prusis, P.; Wikberg, J. E. S. Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinformatics* **2008**, *9*, 181.
(48) Muzammil, S.; Ross, P.; Freire, E. A major role for a set of non-active site mutations in the development of HIV-1 protease drug resistance. *Biochemistry* **2003**, *42*, 631–638.
(49) Rhee, S.-Y.; Fessel, W. J.; Zolopa, A. R.; Hurley, L.; Liu, T.; Taylor, J.; Nguyen, D. P.; Slome, S.; Klein, D.; Horberg, M.; Flamm, J.; Follansbee, S.; Schapiro, J. M.; Shafer, R. HIV-1 protease and reverse-transcriptase mutations: correlations with antiretroviral therapy in subtype B isolates and implications for drug-resistance surveillance. *J. Infect. Dis.* **2005**, *192*, 456–465.

CI1001394