Application of the PharmPrint Methodology to Two Protein Kinases

Felix Deanda*,† and Eugene L. Stewart*,‡

Computational, Analytical, and Structural Sciences, GlaxoSmithKline, Five Moore Drive, Research Triangle Park, North Carolina 27709

Received March 25, 2004

The PharmPrint methodology developed by McGregor and Muskal[1,2] was used to construct quantitative structure−activity relationship (QSAR) models for the prediction of cyclin-dependent kinase-2 (CDK2) and vascular endothelial growth factor receptor-2 (VEGFR2) inhibition. The QSAR models were constructed based on a binary description of biological activity—a value of zero for inactive and one for active compounds. Subsets of "active" kinase inhibitors (that is, inhibitors with $pIC_{50} \geq 6.0$) along with a subset of MDDR[3] compounds serving as the recommended set of inactive compounds were used for model development. The predicted activities for the training set compounds were in excellent agreement with the assigned binary activities with greater than 92% of the compounds correctly classified. However, when the QSAR models were applied to the subsets of "inactive" kinase inhibitors (that is, inhibitors with $pIC_{50} < 6.0$), greater than 67% were incorrectly predicted to be active. Identical results were obtained with our CDK2 and VEGFR2 validation sets, where the majority of the inactive kinase inhibitors were predicted to be active. In efforts to improve the predictive performance of the QSAR models, simple, but important modifications were made to the PharmPrint methodology. On the basis of these modifications, a second set of QSAR models was constructed and applied to our validation sets to assess their predictive performance. Significant improvements were seen with the modified version of PharmPrint over the original. The results from both versions of PharmPrint are compared and discussed.

INTRODUCTION

With the advent of combinatorial chemistry and high-throughput screening technologies, the pharmaceutical and biotechnology industries have at their disposal the tools necessary to synthesize and screen potentially large numbers of compounds. To meet the assumed demands of high-throughput screening, medicinal chemists are designing compound collections from virtual libraries composed of literally millions of compounds. There are also many other sources of compounds including those available for purchase from commercial vendors. Nevertheless, for resource-related reasons, synthesis, or purchase of vast numbers of compounds for biological screening is not practical or desirable. Consequently, an accurate and efficient computational approach is needed for use as a virtual high-throughput screening tool to reduce large sets of compounds to smaller subsets of pharmaceutical interest. Quantitative structure−activity relationship (QSAR) methods represent an attractive solution to this problem given that there are many such applications that are less computation-intensive and less time-consuming compared to other computer-assisted drug design (CADD) methodologies such as ligand docking.[4−8] To this end, we have recently evaluated the PharmPrint methodology as a virtual screening tool for use with protein kinases.

McGregor and Muskal have published details regarding the PharmPrint fingerprint and its use as a QSAR descriptor.[1] Briefly, the PharmPrint fingerprint is a binary bit-string, wherein each bit indicates either the absence or presence of one of 10 549 three-point pharmacophores. The three-point pharmacophores were constructed by enumerating seven pharmacophoric features along with six distance ranges. Six of the pharmacophoric features represent modes of intermolecular interaction commonly used in describing receptor−ligand interactions. These include atoms or groups of atoms with formal positive or negative charges, aromatic groups, hydrophobic groups, and H-bond donors and acceptors. The seventh pharmacophoric feature, labeled as the "X" group, represents any atom not labeled as one of the aforementioned pharmacophoric types. The set of distance ranges used in the enumeration of the PharmPrint three-point pharmacophores were those previously published by Pickett et al.[9] in their work involving diversity profiling of chemical libraries and databases via 3D pharmacophores.

In developing a PharmPrint QSAR model, the first step involves constructing the PharmPrint fingerprints for a training set of compounds. This is accomplished by a series of substructure searches to identify the pharmacophoric features present in a given compound. Next, multiple conformers are generated to determine the achievable distances between pharmacophores. A limit of 1000 conformers per compound is artificially imposed. Thus, rotatable bonds affecting the largest number of atoms are searched first. Once conformer generation is completed, the resultant three-point pharmacophores are identified and the PharmPrint fingerprint constructed accordingly. Finally, given biological data and using the PharmPrint fingerprints as sets of predictor variables, a QSAR model is constructed by partial-least-squares (PLS) regression, wherein the extracted PLS factors are computed via the NIPALS algorithm.[1]

* Corresponding author.
† Telephone: (919) 483-9482; fax: (919) 483-6053; e-mail: Felix.G.Deanda@gsk.com.
‡ Telephone: (919) 483-0152; fax: (919) 315-0430; e-mail: Eugene.L.Stewart@gsk.com.

10.1021/ci0498968 CCC: $27.50
Published on Web 06/24/2004

The PharmPrint package is comprised of a suite of programs, which enable not only the construction of PharmPrint fingerprints but also the development of QSAR models and the application of those models in predicting the biological activities of compounds yet to be synthesized or screened. McGregor and Muskal developed two versions of a PLS algorithm for QSAR model development. The algorithm one uses will depend on the type of biological data that are available. One version constructs QSAR models by fitting experimentally measured biological activities (e.g., $p\mathrm{IC}_{50}$ or $p\mathrm{EC}_{50}$ values) to a set of PharmPrint fingerprints; a second version uses binary activity data (i.e., a value of zero for inactive and one for active compounds) as the dependent variable in the PLS regression.[1] Although the former was preferred, McGregor and Muskal provided only the latter version of their PLS program (henceforth, referred to as the PharmPrint/PLS algorithm). Included with the PharmPrint package was a subset of MDDR compounds to be used as the recommended set of inactive compounds with the PharmPrint/PLS algorithm.

PharmPrint QSAR models were constructed for predicting cyclin-dependent kinase-2 (CDK2) and vascular endothelial growth factor receptor-2 (VEGFR2) inhibition based on a binary description of biological activity. As required, our training sets were divided into subsets of active and inactive kinase inhibitors. The sets of inactive kinase inhibitors were then replaced with the MDDR subset and the QSAR models constructed. As might be expected, when simply fitting experimental data to calculated molecular descriptors, the predicted binary activities for the training set compounds were in excellent agreement with the observed activities. However, when the QSAR models were applied to our subsets of inactive kinase inhibitors as well as our validation sets, the results were disappointing but certainly not unexpected. Over two-thirds of the inactive kinase inhibitors were incorrectly predicted to be active. In efforts to improve the predictive performance of the PharmPrint methodology, the MDDR subset was eliminated altogether and replaced with the known inactive kinase inhibitors. To fully implement this change, however, required that the PharmPrint/PLS algorithm be abandoned in favor of the SAS/PLS procedure from the SAS system.[10] In this report, we detail these simple, but important, modifications. Additionally, results from the original and in-house versions of the PharmPrint methodology are summarized and compared.

METHODOLOGY

Protein Kinase Data Sets. Experimental $p\mathrm{IC}_{50}$ data for CDK2 and VEGFR2 inhibitors were collected from our corporate database for the purposes of constructing and subsequently validating PharmPrint QSAR models. CDK2 and VEGFR2 were selected as representatives of the serine/threonine and tyrosine protein kinase families, respectively, primarily due to the large volume of biological and chemical data that were available. The CDK2 data set was comprised of 2254 inhibitors with $p\mathrm{IC}_{50}$ values ranging from 4.0 to 8.81. In total, 83 different structural classes of compounds were represented in the data set as determined by an in-house chemical classification scheme. The VEGFR2 data set was comprised of 2306 inhibitors representing 85 structural classes with $p\mathrm{IC}_{50}$ values ranging from 3.74 to 9.44.

To assess the predictive abilities of the QSAR models resulting from both the original and in-house versions of PharmPrint, 20% of the compounds from each data set were randomly selected and set aside for use as validation sets. The CDK2 validation set was comprised of 451 compounds with $p\mathrm{IC}_{50}$ values ranging from 4.02 to 8.73. The remaining subset of 1803 inhibitors was used as the training set for constructing QSAR models for CDK2 inhibition. The VEGFR2 validation set included 461 compounds with $p\mathrm{IC}_{50}$ values ranging from 3.92 to 9.36. The remaining 1845 inhibitors were used to construct QSAR models for VEGFR2 inhibition.

The PharmPrint/PLS algorithm requires as input two sets of compounds, a set of "actives" and a set of "inactives". Although compounds at the extreme high and low ends of the $p\mathrm{IC}_{50}$ range are generally regarded as active and inactive, respectively, such classification for compounds with modest $p\mathrm{IC}_{50}$ values is mostly subjective. Nevertheless, we attempted to establish a minimum $p\mathrm{IC}_{50}$ value for active compounds, $p\mathrm{IC}_{50}^{\mathrm{min}}$, to divide our training sets into subsets of actives and inactives. While careful consideration was given to each protein kinase, in the end, the choice of value for $p\mathrm{IC}_{50}^{\mathrm{min}}$ was simply arbitrary. For both training sets, a $p\mathrm{IC}_{50}^{\mathrm{min}}$ value of 6.0 was chosen. Of the 1803 CDK2 inhibitors, 750 compounds had values of $p\mathrm{IC}_{50} \geq 6.0$ and were classified as active. The remaining 1053 compounds were classified as inactive. The 1845 VEGFR2 inhibitors were divided in similar fashion with 965 compounds classified as active and 880 compounds classified as inactive.

As indicated, along with their suite of programs, McGregor and Muskal provided a "background" set of compounds to be used with the PharmPrint/PLS algorithm as the recommended set of inactives. The data set was comprised of 11 699 MDDR[3] compounds and included registry numbers and PharmPrint fingerprints. The criteria for inclusion of these compounds into the background set are assumed to be identical to those previously published for the MDDR9104 subset.[1,2] As appropriate, known kinase inhibitors were eliminated from the background set. A total of 346 kinase inhibitors were identified on the basis of their MDDR activity class (the class designations were obtained from our in-house MDDR database) and subsequently eliminated to yield a final set of 11 353 MDDR compounds.

Given that the MDDR database represents a knowledge base of drug and druglike compounds, the background set may appear to be a reasonable alternative to a subset of known inactive compounds. Moreover, the background set may introduce greater structural diversity compared to the subset of inactive compounds, thus potentially leading to a more general QSAR model. Nevertheless, we must assume that the MDDR compounds are inactive against our protein kinases, clearly a questionable assumption for any protein target and one recognized by McGregor and Muskal. Indeed, the term "background" was used as a means to distinguish true inactive compounds, as properly identified through biological screening, from compounds assumed to be inactive against a biological target of interest.[1] In efforts to alleviate the uncertainties introduced into the resultant QSAR models by such an assumption, the PharmPrint/PLS algorithm utilizes the Tanimoto coefficient (calculated on the basis of PharmPrint fingerprints) to identify MDDR compounds that

are "pharmacophorically" similar to those in the training set of actives. These MDDR compounds are then eliminated prior to QSAR model development since they may possess biological activity against the target of interest. The user is required to specify, as an input parameter to the PharmPrint/PLS algorithm, a cutoff value for the Tanimoto coefficient. Inactive compounds, which either meet or exceed the user-specified value when computed against any active compound, are eliminated. In the present work, a conservative value of 0.70 was used.

Modifications to PharmPrint. From a practical and theoretical standpoint, we recognize that compounds assumed to be inactive against a protein target of interest should not be used for QSAR model development. Although the MDDR background set may possess greater structural diversity, consider the fact that each of our kinase data sets is not simply a congeneric series of compounds but a diverse set representing numerous in-house structural classes. Given the questionable validity of the assumption made with regard to the background set, the uncertainties introduced into the resultant QSAR models by such an assumption and the fact that we actually have known inactives, we replaced the MDDR subset with our own subsets of inactive kinase inhibitors and proceeded to construct a second set of QSAR models.

The training subsets of active and inactive VEGFR2 inhibitors were provided as input to the PharmPrint/PLS algorithm. Upon execution, the program reported that the number of active compounds had to be less than or equal to the number of inactives. This requirement was unexpected and one that the VEGFR2 training subsets did not satisfy. Unless the value of $p\mathrm{IC}_{50}^{\mathrm{min}}$ was raised or roughly 9% of the active compounds eliminated, a QSAR model could not be constructed. Although the CDK2 training subsets of active and inactive inhibitors did satisfy the above requirement, we discovered that the PharmPrint/PLS algorithm had eliminated 303 inactive compounds prior to QSAR model development. This occurred despite the fact that no inactive compound was found to have a fingerprint identical to that of any active compound. Given that the identities of the eliminated compounds were not reported, we could not investigate this matter further, but were left to conclude that the kinase inhibitors were simply eliminated to make the number of inactive compounds equal the number of actives.

In effect, the PharmPrint/PLS algorithm imposes the requirement that the number of active compounds in a training set equals the number of inactives. Rather than have our kinase data sets conform to this requirement and be modified for no apparent reason, we opted to replace the PharmPrint/PLS algorithm with the SAS/PLS procedure from the SAS system for QSAR model development. The SVD algorithm was selected for latent factor extraction. Although less efficient than the NIPALS algorithm, it is the most accurate among the SAS/PLS factor extraction methods.[10] Split cross-validation was performed to identify the optimal number of PLS components to include in each QSAR model. The statistic used by the SAS/PLS procedure in judging the number of latent factors to include was the root mean PRESS.

Since the PharmPrint/PLS algorithm was no longer being used, we also chose to replace the binary activities with the experimental $p\mathrm{IC}_{50}$ values, thus eliminating the need for an arbitrary $p\mathrm{IC}_{50}^{\mathrm{min}}$ value to divide the training sets into subsets of active and inactive compounds. To import the PharmPrint fingerprints for a training set of compounds into the SAS/PLS procedure, a program was written to translate the binary fingerprint into ASCII text. This resulted in very large data files compared to those containing the more compact binary fingerprint. To help reduce the size of the ASCII files, the PharmPrint fingerprint was enriched for each training set by eliminating three-point pharmacophores absent from the given data set. Clearly, if a three-point pharmacophore is absent from a set of compounds, it can be safely eliminated without affecting the PLS regression model. Three-point pharmacophores present in all training set compounds were also considered for elimination, given that these descriptors are essentially information-poor (i.e., these descriptors do not provide a basis for modeling the different biological activities seen among the kinase inhibitors). However, no such three-point pharmacophores were identified in either training set.

RESULTS AND DISCUSSION

PharmPrint/PLS QSAR Models. For this study, two training sets of compounds were assembled for the purpose of constructing PharmPrint QSAR models. One set was comprised of 1803 CDK2 inhibitors; the second set contained a total of 1845 VEGFR2 inhibitors. As stated previously, given that the PharmPrint/ PLS algorithm requires as input a set of "actives" and a set of "inactives", each training set was divided into two subsets. Kinase inhibitors with experimental values of $p\mathrm{IC}_{50} \geq 6.0$ were classified as active and their corresponding biological activities reassigned a (binary) value of one. Although the remaining kinase inhibitors were classified as inactive, they were not used for QSAR model development. Instead, the background set of MDDR compounds was used as the recommended set of inactives. Each MDDR compound was assigned a (binary) biological activity of zero, thus assuming inactivity against the protein kinases of interest. To alleviate the uncertainties associated with such an assumption, MDDR compounds yielding a Tanimoto coefficient ≥ 0.70 when computed against any active compound were eliminated by the PharmPrint/PLS algorithm prior to QSAR model development.

A QSAR model was constructed from the binary biological activities for the sets of active CDK2 inhibitors and MDDR compounds. The binary activities were fitted against the PharmPrint fingerprints, resulting in a six-component PLS model. The PharmPrint/PLS algorithm reported that the QSAR model was able to account for 74% of the variance in the binary biological data with a root-mean-squared error of 0.255. The predicted activities for the active CDK2 inhibitors ranged from 0.159 to 1.358 with an average of 0.870 and a standard deviation of 0.210. For the MDDR subset, the predicted activities ranged from -0.606 to 1.149 with an average of 0.141 and a standard deviation of 0.234. To transform these calculated values into predicted binary activities, the PharmPrint/PLS algorithm recommended that 0.50 be used as the minimum value for classifying compounds as active. Thus, compounds with calculated values ≥ 0.50 were to be classified as active, while all others were to be classified as inactive. On the basis of this recommendation, we found that, of the 750 active CDK2 inhibitors, 93.5%

Table 1. PharmPrint/PLS QSAR Results for the Subset of Active and Inactive Compounds in the Kinase Training Sets

	active compounds			inactive compounds		
protein kinase	total	predicted active	predicted inactive	total	predicted inactive	predicted active
CDK2	750	701 (93.5%)	49 (6.50%)	1053	334 (31.7%)	719 (68.3%)
VEGFR2	965	923 (95.6%)	42 (4.40%)	880	288 (32.7%)	592 (67.3%)

were correctly predicted to be active (see Table 1). For the MDDR subset, 92.1% of the compounds were correctly predicted to be inactive.

A QSAR model was also constructed from the binary activities for the set of active VEGFR2 inhibitors. Once again, the MDDR background set was used as the training set of inactive compounds. The resultant PharmPrint QSAR model was comprised of six PLS components and accounted for 79% of the variance in the binary biological data with a root-mean-squared error of 0.229. The predicted activities for the VEGFR2 inhibitors ranged from 0.146 to 1.331 with an average of 0.895 and a standard deviation of 0.191. For the MDDR subset, the minimum and maximum predicted activities were −0.789 and 1.255, respectively, with an average of 0.122 and a standard deviation of 0.231. As with the CDK2/MDDR set, the recommended minimum calculated value for classifying a compound as active was 0.50. On the basis of this recommendation, we found that, of the 965 active VEGFR2 inhibitors, 95.6% were correctly predicted to be active (see Table 1). For the MDDR subset, 92.5% of the compounds were correctly predicted to be inactive.

Having derived QSAR models based on the CDK2/MDDR and VEGFR2/MDDR training sets, the models were then applied to the subsets of inactive CDK2 and VEGFR2 inhibitors. Note from Table 1 that a large percentage of these compounds were incorrectly predicted to be active. Of the 1053 inactive CDK2 inhibitors, 68.3% were predicted to be active. Similar results were seen with the 880 inactive VEGFR2 inhibitors, where 67.3% were predicted to be active. These results were not surprising for the following reasons. First, consider that structural similarities exist among the active and inactive kinase inhibitors of a given chemical class. Indeed, there are countless examples in the scientific literature where nearly identical compounds have been shown to have significantly different biological activities against a target of interest. To encode the structural differences that lead to different biological activities requires that a QSAR model be properly trained on known active and inactive compounds. Since we have replaced our known inactive kinase inhibitors with the MDDR subset, the QSAR models cannot be expected to effectively distinguish active kinase inhibitors from inactive. Second, there is the questionable assumption made with regard to the biological activities of the MDDR compounds and the uncertainties that this introduces into the resultant QSAR models.

Given these two observations, it is reasonable to expect that these QSAR models will yield high false positive rates when applied as virtual screens against chemical databases or virtual libraries. This would then require that we assemble larger compound sets for biological screening than should be necessary. To address the first issue, known inactive kinase inhibitors need to be included as part of the training set of inactive compounds. To address the second issue, either the MDDR compounds must be screened against the protein kinases of interest to establish their biological activities or they must simply be eliminated from the training set.

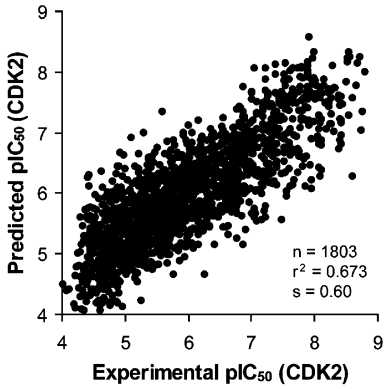

Figure 1. Predicted vs experimental pIC_{50} values for the training set of CDK2 inhibitors. The variables n, r^2, and s represent the number of compounds, the coefficient of determination, and the standard error of prediction, respectively.

SAS/PLS-Derived QSAR Models. Given the poor results seen with the subsets of inactive CDK2 and VEGFR2 inhibitors, we opted to modify the PharmPrint methodology in efforts to improve the predictive performance of the QSAR models. Two rational arguments have been made to support replacing the MDDR background set with known inactive kinase inhibitors. However, to fully implement this change, the PharmPrint/PLS limitations discussed earlier had to be circumvented. This required that the PharmPrint/PLS algorithm be abandoned in favor of a more versatile statistical tool. The SAS/PLS procedure from the SAS system was chosen to replace the PharmPrint/PLS algorithm for QSAR model development. Since the PharmPrint/PLS algorithm was no longer being used, the binary activity data were replaced with the experimental pIC_{50} values, thus eliminating the arbitrary pIC_{50}^{min} value that unnecessarily divided our training sets into subsets of active and inactive compounds.

A SAS/PLS-derived QSAR model was constructed for each training set of kinase inhibitors by fitting the experimental pIC_{50} values against the enriched PharmPrint fingerprints. Recall that the PharmPrint fingerprint was modified for each training set by eliminating three-point pharmacophores absent from all compounds in the given data set. There were 3882 three-point pharmacophores absent from the training set of CDK2 inhibitors. Thus, the enriched PharmPrint fingerprint for this data set included only 6667 three-point pharmacophores. For the VEGFR2 training set, 3519 three-point pharmacophores were identified as absent from all compounds in the data set. Thus, the enriched PharmPrint fingerprint for this data set included only 7030 three-point pharmacophores.

Illustrated in Figure 1 is a plot of predicted versus experimental pIC_{50} values for the training set of CDK2 inhibitors. The SAS/PLS-derived QSAR model was comprised of 13 PLS components with a cross-validated r^2 value of 0.568 and a root mean PRESS of 0.836. The non-cross-

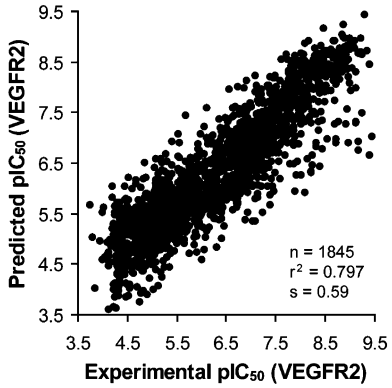

Figure 2. Predicted vs experimental $p\mathrm{IC}_{50}$ values for the training set of VEGFR2 inhibitors. For the definitions of the variables n, r^2, and s see Figure 1.

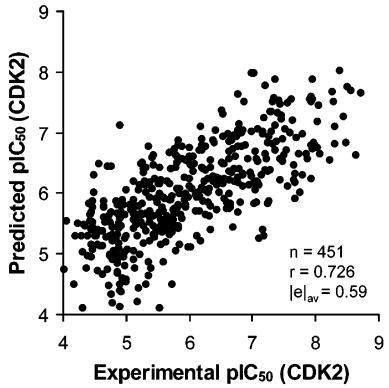

Figure 3. Predicted vs experimental $p\mathrm{IC}_{50}$ values for the validation set of CDK2 inhibitors. The variables n, r, and $|e|_{av}$ represent the number of compounds, the correlation coefficient, and the average unsigned error, respectively.

validated r^2 value was equal to 0.673, and the standard error of prediction was 0.60 log unit. The average unsigned error between predicted and experimental $p\mathrm{IC}_{50}$ values was 0.48 log unit. A plot of predicted versus experimental $p\mathrm{IC}_{50}$ values for the training set of VEGFR2 inhibitors is illustrated in Figure 2. The SAS/PLS-derived QSAR model for this set was comprised of 14 PLS components with a cross-validated r^2 value of 0.695 and a root mean PRESS of 0.722. The non-cross-validated r^2 value was equal to 0.797, and the standard error of prediction was 0.59 log unit. The average unsigned error between predicted and experimental $p\mathrm{IC}_{50}$ values was 0.45 log unit.

The experimental error associated with each training set has been estimated to be less than 0.50 log unit. One potential problem with any regression model is that it may have overfit the experimental data. On the other hand, if the standard error of prediction is much larger than the experimental error, then the model may not be very useful. Given our assessment of the experimental error associated with the biological data, the SAS/PLS-derived QSAR models do not overfit the data. Moreover, the values for the standard error of prediction are reasonable given the correlation observed between predicted and experimental values.

Protein Kinase Validation Sets. The CDK2 and VEGFR2 validation sets were used to assess the predictive abilities of the PharmPrint/PLS and SAS/PLS-derived QSAR models. In this section, the results obtained from the SAS/PLS-derived models are summarized first followed by the results obtained from the PharmPrint/PLS models in comparison to those from the SAS/PLS-derived models.

A plot of predicted versus experimental $p\mathrm{IC}_{50}$ values for the test set of CDK2 inhibitors is illustrated in Figure 3. Note that there is reasonable agreement between both sets of values as evidenced by a correlation coefficient of 0.726. The average unsigned error between predicted and experimental $p\mathrm{IC}_{50}$ values was 0.59 log unit. Illustrated in Figure 4 is a plot of predicted versus experimental $p\mathrm{IC}_{50}$ values for the test set of VEGFR2 inhibitors. Here again, there is reasonable agreement between predicted and experimental $p\mathrm{IC}_{50}$ values with a correlation coefficient of 0.816. The average unsigned error between predicted and experimental $p\mathrm{IC}_{50}$ values for this data set was 0.60 log unit.

To compare results obtained from both versions of PharmPrint, the validation sets were divided into subsets of active and inactive compounds using a $p\mathrm{IC}_{50}^{\mathrm{min}}$ value of 6.0.

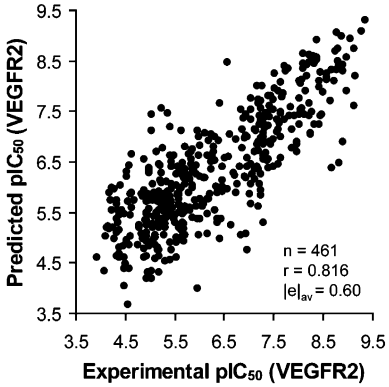

Figure 4. Predicted vs experimental $p\mathrm{IC}_{50}$ values for the validation set of VEGFR2 inhibitors. For the definitions of the variables n, r, and $|e|_{av}$ see Figure 3.

Table 2. Comparison of Correct and Incorrect Predictions by the PharmPrint/PLS and SAS/PLS-Derived QSAR Models for the CDK2 Validation Set

	compnds predicted to be active			compnds predicted to be inactive		
QSAR model	total	active	inactive	total	inactive	active
PharmPrint/PLS	354	186 (52.5%) (92.5%)[a]	168 (47.5%) (67.2%)[b]	97	82 (84.5%) (32.8%)[b]	15 (15.5%) (7.50%)[a]
SAS/PLS-derived	219	168 (76.7%) (83.6%)[a]	51 (23.3%) (20.4%)[b]	232	199 (85.8%) (79.6%)[b]	33 (14.2%) (16.4%)[a]

[a] Percentage of the 201 active compounds ($p\mathrm{IC}_{50} \geq 6.0$). [b] Percentage of the 250 inactive compounds ($p\mathrm{IC}_{50} < 6.0$).

Of the 451 CDK2 inhibitors, 201 compounds had $p\mathrm{IC}_{50}$ values ≥ 6.0 and were classified as active. The remaining 250 compounds were classified as inactive. The 461 VEGFR2 inhibitors were divided in a similar fashion with 217 compounds classified as active and 244 compounds classified as inactive. Comparison of results between the PharmPrint/PLS and SAS/PLS-derived QSAR models in terms of percentages of correct and incorrect predictions are summarized in Tables 2 and 3 for CDK2 and VEGFR2, respectively.

Note from Table 2 that the PharmPrint/PLS QSAR model predicted 354 CDK2 inhibitors to be active. These compounds represent roughly 78% of the inhibitors in the test set. Considering that less than 45% of the CDK2 inhibitors

Table 3. Comparison of Correct and Incorrect Predictions by the PharmPrint/PLS and SAS/PLS-Derived QSAR Models for the VEGFR2 Validation Set

	compnds predicted to be active			compnds predicted to be inactive		
QSAR model	total	active	inactive	total	inactive	active
PharmPrint/PLS	393	205 (52.2%) (94.5%)[a]	188 (47.8%) (77.0%)[b]	68	56 (82.4%) (23.0%)[b]	12 (17.6%) (5.50%)[a]
SAS/PLS-derived	244	188 (77.0%) (86.6%)[a]	56 (23.0%) (23.0%)[b]	217	188 (86.6%) (77.0%)[b]	29 (13.4%) (13.4%)[a]

[a] Percentage of the 217 active compounds ($p\text{IC}_{50} \geq 6.0$). [b] Percentage of the 244 inactive compounds ($p\text{IC}_{50} < 6.0$).

were classified as active, it was not surprising to find that only 52.5% of the compounds predicted to be active were actually active while the remaining 47.5% were inactive. Note the significantly smaller number of compounds predicted to be active by the SAS/PLS-derived QSAR model. Of the 219 compounds predicted to be active, 76.7% were active; the remaining 23.3% were inactive. Clearly, significant improvement in the ratio of active to inactive compounds for the set of predicted actives was achieved. While the PharmPrint/PLS QSAR model did correctly classify 92.5% of the active compounds as active compared to 83.6% by the SAS/PLS-derived model, it did incorrectly classify 67.2% of the inactive compounds as active compared to only 20.4% by the SAS/PLS-derived model. Of the 97 compounds predicted to be inactive by the PharmPrint/PLS QSAR model, 84.5% were inactive; the remaining 15.5% were active compounds. For the SAS/PLS-derived QSAR model, similar positive percentages were seen where 85.8% of the 232 compounds predicted to be inactive were inactive and only 14.2% were active. Note, however, that the SAS/ PLS-derived QSAR model did identify 79.6% of the inactive compounds compared to only 32.8% by the PharmPrint/PLS model.

Listed in Table 3 are the results from the PharmPrint/PLS and SAS/PLS-derived QSAR models for the VEGFR2 test set. These results are comparable to those seen with the CDK2 test set. Note that roughly 85% of the VEGFR2 inhibitors were predicted to be active by the PharmPrint/PLS QSAR model. Given that roughly 47% of the inhibitors were classified as active, the QSAR model demonstrates little discriminating power between known active and inactive compounds. Of the 393 compounds predicted to be active, 52.2% were actually active; the remaining 47.8% were inactive. Again, significant improvement in the ratio of active to inactive compounds for the set of predicted actives by the SAS/PLS-derived QSAR model was achieved. Of the 244 compounds predicted to be active, 77% were active; the remaining 23% were inactive. As seen with the CDK2 test set, although the PharmPrint/PLS model did correctly classify 94.5% of the active compounds as active compared to 86.6% by the SAS/PLS-derived model, it did incorrectly classify 77% of the inactive compounds as active compared to only 23% by the SAS/PLS-derived model. Only 68 compounds were predicted to be inactive by the PharmPrint/PLS QSAR model of which 82.4% were inactive and 17.6% active. For the SAS/PLS-derived QSAR model, 217 compounds were predicted to be inactive of which 86.6% were inactive and only 13.4% active. Note that the SAS/PLS-derived model did correctly classify 77% of the inactive compounds as inactive compared to only 23% by the PharmPrint/PLS model.

CONCLUSIONS

The PharmPrint methodology of McGregor and Muskal was applied to CDK2 and VEGFR2 inhibitors to assess its utility as a virtual screening tool for protein kinases. PharmPrint QSAR models were constructed on the basis of binary biological activities for subsets of active CDK2 and VEGFR2 inhibitors and assumed inactive MDDR compounds. The predicted activities for the training set compounds were in excellent agreement with the assigned binary activities. Greater than 92% of the active and inactive compounds were correctly classified. However, when the QSAR models were applied to the inactive CDK2 and VEGFR2 inhibitors, the results were disappointing. Over two-thirds of these compounds were incorrectly predicted to be active. These results were not surprising considering the questionable assumption made with regard to the biological activities of the MDDR compounds and the negative impact of such an assumption on the resultant QSAR models. In addition, and perhaps more importantly, structural data relevant to biological activity were discarded when the MDDR compounds were used for QSAR model development in place of known inactive kinase inhibitors.

Modifications were made to the PharmPrint methodology in efforts to improve the predictive abilities of the QSAR models. The most significant change involved replacing the MDDR background set with known kinase inhibitors. To fully implement this change required that we abandon the PharmPrint/PLS algorithm in favor of the SAS/PLS procedure. Since the PharmPrint/PLS algorithm was no longer being used, the binary activity data were also replaced with experimental $p\text{IC}_{50}$ values, eliminating the need to arbitrarily divide our training sets into subsets of active and inactive compounds. Finally, since it was necessary to translate the binary bit-string into ASCII format for import into the SAS/PLS procedure, to help reduce the size of the files, the PharmPrint fingerprint was enriched by eliminating three-point pharmacophores absent from all compounds in a given training set. Note that the modifications only involved the PharmPrint methodology and not the kinase training sets. As such, these changes should be generally applicable to any medicinal class of compounds.

Having completed all modifications, the in-house version of the PharmPrint methodology was applied to the training sets of CDK2 and VEGFR2 inhibitors. A 13-component QSAR model was constructed from the set of CDK2 inhibitors. The model was able to account for 67.3% of the variance in the experimental data with a standard error of prediction of 0.60 log unit. A 14-component QSAR model was also constructed from the set of VEGFR2 inhibitors. This model was able to account for 79.7% of the variance in the experimental data with a standard error of prediction of 0.59 log unit. The QSAR models were then applied to the validation sets of CDK2 and VEGFR2 inhibitors. For CDK2, the agreement between predicted and experimental values was reasonable with a correlation coefficient of 0.726 and an average unsigned error of 0.59 log unit. The agreement between predicted and experimental values for

VEGFR2 was also reasonable with a correlation coefficient of 0.816 and an average unsigned error of 0.60 log unit. For comparison purposes only, the validation sets were divided into subsets of active and inactive compounds. As a result of the modifications, the percentages of correct and incorrect predictions were significantly improved with the SAS/PLS-derived QSAR models as compared to the PharmPrint/PLS models.

REFERENCES AND NOTES

(1) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 1. Application to QSAR and Focused Library Design. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 569−574.

(2) McGregor, M. J.; Muskal, S. M. Pharmacophore Fingerprinting. 2. Application to Primary Library Design. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 117−125.

(3) *MDL Drug Data Report (MDDR)*; MDL Information Systems, Inc.: San Leandro, CA, 2000.

(4) Tong, W.; Lowis, D. R.; Perkins, R.; Chen, Y.; Welsh, W. J.; Goddette, D. W.; Heritage, T. W.; Sheehan, D. M. Evaluation of Quantitative Structure−Activity Relationship Methods for Large-Scale Prediction of Chemicals Binding to the Estrogen Receptor. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 669−677.

(5) Martin, Y. C. 3D QSAR: Current State, Scope, and Limitations. In *Perspectives in Drug Discovery and Design*; Kubinyi, H., Folkers, G., Martin, Y. C., Eds.; Kluwer Academic Publishers: London, Great Britain, 1998; Vols. 12/13/14, pp 3−23.

(6) Ivanciuc, O.; Taraviras, S. L.; Bass, D. Quasi-orthogonal Basis Sets of Molecular Graph Descriptors as a Chemical Diversity Measure. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 126−134.

(7) Golbraikh, A.; Shen, M.; Xiao, Z.; Xiao, Y.; Lee, K.; Tropsha, A. Rational Selection of Training and Test Sets for the Development of Validated QSAR Models. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 241−253.

(8) Stanton, D. T. On the Physical Interpretation of QSAR Models. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1423−1433.

(9) Pickett, S. D.; Mason, J. S.; McLay, I. M. Diversity Profiling and Design Using 3D Pharmacophores: Pharmacophore-Derived Queries (PDQ). *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 1214−1223.

(10) *The SAS System for Windows*, Version 8.01; SAS Institute, Inc.: Cary, NC, 2000.

CI0498968