Proteochemometric Modeling of Drug Resistance over the Mutational Space for Multiple HIV Protease Variants and Multiple Protease Inhibitors

Maris Lapins and Jarl E. S. Wikberg*

Department of Pharmaceutical Pharmacology, Uppsala University, SE-751 24 Uppsala, Sweden

Received December 15, 2008

The main therapeutic targets in HIV are its protease and reverse transcriptase. A major problem in treatment of HIV is the ability of the virus to develop drug resistance by accumulating mutations in its targets. Acquiring detailed understanding of the molecular mechanisms for the interactions of drugs with mutated variants of the HIV virus is mandatory to be able to design inhibitors that can evade the resistance. Here we have used proteochemometric modeling to simultaneously analyze the interactions of 21 protease inhibitors with 72 unique protease variants. Inhibition data (pK_i) were correlated to descriptions of chemical and structural properties of the inhibitors and proteases. The proteochemometric model obtained showed excellent fit and predictive ability (R^2=0.92, Q^2=0.83, Q^2_{inh}=0.78) and provided quantitative assessments for the contribution of each mutation and their combinations to the decrease in inhibitor activity, both for the whole compounds series as well as for individual compounds. The model revealed the most deleterious mutations in the protease to be D30N, V32I, G48V, I50V, I54V, V82A, I84V, and L90M. The model was further used to identify molecular properties of chemical compounds that are important for their inhibition of multimutated protease variants. Our results give directions how to design novel improved inhibitors.

INTRODUCTION

HIV protease is a small homodimeric protein composed of two identical 99-residue chains. The enzyme ensures cleavage of the viral precursor Gag-Pol polyprotein and is essential for the maturation of the virus into infectious particles. Along with reverse transcriptase, it is a target for highly active antiretroviral therapy (HAART); nine of the 30 currently used anti-HIV drugs are inhibitors of the protease.

Since its introduction in 1996, HAART has resulted in markedly prolonged patient survival. However, the high replication rate and the error prone mechanism of retroviral transcription permit HIV to mutate and develop resistance, thus allowing it to escape from the antiviral therapy. Resistance emerges rapidly when protease and reverse transcriptase inhibitor combinations are administered at inadequate doses or at regimens where the viral replication is only partially suppressed. However, even when the levels of HIV are oppressed below the detectable level in plasma, the virus remains in cellular reservoirs where it slowly evolves.[1]

Resistance to protease inhibitors develops primarily by accumulating mutations in the enzyme itself and to a minor extent also by mutations in the cleavage sites of the Gag-Pol polyprotein.[2] Data from phenotypic drug susceptibility assays have been collected for several thousand viral isolates,[3] and phenotype-genotype correlations show that resistance inducing mutations occur both in the active site of the protease and outside it (see 4 and references therein). In the past distantly located mutations were generally regarded as "compensatory" and supposed to counteract the negative effects of the active-site mutations on the conformation and dimerization of the protease and its ability to cleave substrates. However, more recent studies suggest that such mutations may also diminish the drugs' ability to bind the protease by indirectly altering the geometry of the active site; hence the influence of distant mutations should also be accounted for in the design of inhibitors.[5] The two mechanisms cannot be distinguished by the analysis of the results of phenotypic assays alone and adds to the difficulties to design drugs for mutated forms of the HIV protease.

As different drugs show different inhibition profiles, selection of a proper regimen based on viral genotype can to some extent surmount the resistance. However, certain mutations are found to reduce susceptibility to all of the inhibitors in current clinical use.[4] This phenomenon of cross-resistance prompts the need for novel more adaptive agents that can oppress the whole mutational space of the HIV protease.

A widely applied computational method for the optimization of chemical compounds' interaction with a target protein is quantitative structure−activity relationship (QSAR) modeling. Quite many studies have applied QSAR to HIV protease-inhibitor interactions.[6−9] However, only a few address inhibitor interactions with protease mutants. One was published by Avram et al.[10] who created separate models for inhibitory activity (pK_i) of 14 cyclic urea derivatives for the three frequently occurring protease mutants V82A, V82I, and V82F. Another recent study exploited qualitative data (i.e., high/low activity) for nine inhibitors tested against mutants V82A, V82F, and I84V.[11] Separate models were created for each protease variant and used to predict their susceptibility to over 40 other compounds.

Separate modeling for one mutant at a time shows several drawbacks, however. First, each model of the above studies

* Corresponding author e-mail: Jarl.Wikberg@farmbio.uu.se.

10.1021/ci800453k CCC: $40.75
Published on Web 04/27/2009

exploited quite small amounts of data, which limits model interpretations severely. Second, a major limitation of the separate QSAR modeling approach is that the obtained models cannot be used to predict inhibition of proteases which comprise numerous mutations in many different and even new combinations. To cope with these limitations we have applied proteochemometrics, an approach that we developed some time ago to study the mechanisms for molecular recognition of groups of related proteins.[12] Proteochemometric models are based on experimentally determined interaction data for proteins interacting with series of ligands (e.g., organic compounds, peptide substrates, etc.). These data are simultaneously correlated to physicochemical and/or structural variations within the two sets of interacting entities and lead to models that reveal properties determining protein−ligand recognition which can be used to guide the design of novel entities with improved complementarity. We have previously successfully applied proteochemometrics to model ligand interactions with various classes of G-protein coupled receptors.[13−15] The aim of this study was to use proteochemometrics to analyze inhibitor interactions with a pool of multiple mutated HIV proteases.

METHODS

Data Set. Inhibition constants (pK_i) of 21 chemical compounds for the wild-type and 71 mutated HIV protease variants (totally 386 unique inhibitor-protease combinations) were collected from the literature (see the Supporting Information).[5,16−36] The sequences of mutated proteases differed from the wild-type sequence by from one to twelve (on the average 4.4) amino acid substitutions. The compound series included the FDA approved drugs amprenavir, atazanavir, darunavir, indinavir, lopinavir, nelfinavir, ritonavir, and saquinavir and other inhibitors that had reported pK_i data for the wild-type protease and at least five protease mutants; the PubChem CID numbers of these compounds were as follows: 60927, 64999, 72404, 200104, 445303, 445305, 445308, 449114, 462367, 470657, 475872, 475876, and 501464. All compounds were potent inhibitors for the wild-type protease, their inhibition constants ranging from low picomolar up to 19 nM, while the inhibition constants for mutated proteases covered a range of more than six logarithmic units.

For part of the inhibitor-protease combinations pK_i values were available from two or more sources. These overlapping data allowed us to judge the consistency of the measurements; the root-mean-square deviation from the mean being 0.56 logarithmic units. However, in a few cases more than a 100-fold discrepancy was found between two sources, indicating that differences in assaying conditions may have a large influence on the measurement results, something which accordingly should be accounted for in the proteochemometric modeling (*vide infra*).

Description of Proteases. Of the 99 amino acids in each protease monomer, 40 were found to be mutated in the data set. Of the latter 32 showed a variation of only two amino acids for each position: one corresponding to the wild-type sequence and the other to a mutant variant. For each of these 32 positions a binary descriptor was assigned with a value 0 if the wild-type amino acid was present; otherwise it was set to 1. Each of the remaining eight positions showed a variation of more than two different amino acids. These positions were described by amino acid physicochemical properties encapsulated in the three so-called zz-scales, zz_1-zz_3, derived by Sandberg et al.[37] zz-Scales are obtained by applying principal component analysis[38] to 26 measured and computed physicochemical properties of amino acids and represent essentially hydrophobicity (zz_1), steric properties (zz_2), and polarity (zz_3) of amino acids. In this way, the varying parts of protease sequences were represented by $32 + 8 \times 3 = 56$ descriptors.

Description of Protease Inhibitors. 3D structures of organic compounds were created by the Corina unit of the Tsar 3.3 (Accelrys Inc.) software and optimized by applying semiempirical quantum mechanical calculations using the Vamp unit of Tsar 3.3. Compounds were then characterized by molecular descriptors representing properties that are related to molecular shape, flexibility, and the ability for different types of noncovalent interactions. To this end the following descriptor classes were calculated by use of Dragon 2.1 software (Talete S.r.l.): geometrical descriptors, charge and aromaticity indices, constitutional descriptors, counts of functional groups and atom-centered fragments, empirical descriptors, and molecular properties. Descriptors were checked for mutual correlation, and when two or more descriptors were found to be highly correlated (pairwise $r^2 > 0.9$) only one of them was retained. In this way, 84 molecular descriptors were obtained for the modeling (see the caption for Figure 5 for the descriptors list).[39]

Laboratory Descriptors. As stated above inspection of the inhibition data collected from the 22 literature sources revealed some discrepancies in the reported pK_i values that could result from differences in assay conditions. To normalize for possible systematic deviations in measurements from different laboratories we included one binary indicator variable for each source. A variable was thus assigned the value 1 if the measurement was performed by the given data source; otherwise it was assigned 0.

Protease-Inhibitor and Intraprotease Cross-Terms. Protein−ligand interactions are governed by complex processes that depend on the complementarity of the properties of the interacting entities. In proteochemometrics this can be accounted for by simultaneous description of the two interacting entities by so-called cross-descriptions (also called cross-terms), which is in the simplest case obtained by multiplication of mean centered descriptors of proteins and ligands. Cross-terms can also be computed between protein descriptors as well as between ligand descriptors. The latter terms are required in situations when two different properties show co-operative effects and can, e.g., relate to complex changes within proteins, such as intramolecular interactions and conformational effects.

In order to account for effects of particular mutations on the pK_i of particular inhibitors, we computed cross-terms between protease and inhibitor descriptors, which yielded 4704 (56×84) cross-terms. To account for the eventual co-operation of protease residues in conferring resistance we also introduced cross-terms between protease descriptors, which yielded $(56 \times 55)/2 = 1{,}540$ intraprotease cross-terms.

Preprocessing of Data. Prior to calculation of cross-terms and further modeling all descriptors were mean centered and scaled to unit variance. Moreover, to account for differences in the number of descriptors and cross-terms we applied

block scaling. The scaling factor for the cross-terms was increased stepwise, starting from 0, by 0.05 unit intervals. An optimal model was obtained by increasing the scaling factor until the predictive ability of the model assessed by cross-validation (*vide infra*) reached the maximum.

The dependent variable (pK_i) was also mean centered prior to use in the computations.

Correlation by Partial Least-Squares Projections to Latent Structures. The above-derived descriptors and cross-terms (independent variables comprised in the **X** matrix) were correlated to the pK_i (dependent **y** variable) by using the partial least-squares projections to latent structures (PLS). In PLS, **X** and **y** are simultaneously projected to latent variables (PLS components), with an additional constraint to maximize the covariance between the projections of **X** and **y**.[40] PLS derives a regression equation where regression coefficients reveal the direction and magnitude of the influence of **X**-variables on **y**. Thus, for a model comprising L ligand descriptors, P protease descriptors, and ligand-protease and intraprotease cross-terms, the PLS modeling results can be expressed as follows:

$$y = \bar{y} + \sum_{l=1}^{L} (coeff_l x_l) + \sum_{p=1}^{P} (coeff_p x_p) + \sum_{l=1,p=1}^{L \times P} (coeff_{l,r} x_l x_p) + \sum_{p_1 \neq p_2}^{p \times (p-1)/2} (coeff_{p_1,p_2} x_{p1} x_{p2})$$

Several algorithms have been developed for performing PLS; here we used orthogonalized-PLS[41,42] as implemented in the P software under Bioclipse, Bioclipse-P.[43,44]

Validation of Modeling. Models were thoroughly validated to find the optimal scaling for cross-terms and optimal model complexity (number of PLS components). The goodness-of-fit of the PLS models was characterized by the fraction of explained variance of **y** (R^2). The predictive ability was characterized by the fraction of the predicted **y**-variance (Q^2), assessed by cross-validation with seven randomly formed groups, as previously described.[45] In PLS, the R^2 term increases with each extracted PLS component, while the Q^2 value usually reaches a plateau and declines as the model becomes overfitted. Similarly, high scaling weights for nonlinear terms allows better fit of the model but may decrease its predictive ability.

Since our data set contained replicates (i.e., pK_i values for 49 of 386 inhibitor-protease combinations were available from several sources), cross-validation groups were formed so that replicates were always placed in the same group. Thus, when predictions for a particular drug-protease pair were performed, no data for this pair had been used in the creation of the model.

We also wanted to assess the ability of the proteochemometric model to predict inhibition of the HIV proteases by novel compounds that had not been present in the model in any combination with the mutated proteases. Therefore, we performed cross-validation by excluding one inhibitor at a time and then predicted its pK_i for all protease variants; the results from this modified cross-validation procedure are referred to here as Q^2_{inh}.

Finally, we assessed the relationship between the sparseness of the data set and the predictive performance and robustness of the resulting models. To do this, we divided the data set into four parts and created models using only three, two, or one of these parts. The remaining data were put aside during the whole modeling procedure (i.e., these data did not influence centering and scaling of descriptors, the finding of the optimal scaling for cross-terms, and the optimal PLS model complexity) and were used after creation of the model to evaluate its external predictive ability. Splitting the data set was performed randomly and repeated twenty times. The external predictive ability P^2 and standard deviation of this parameter was calculated for each size of the data set from the twenty splits.

RESULTS AND DISCUSSION

Performance of the Proteochemometric Model. Descriptors of proteases, inhibitors, and protease-inhibitor cross-terms were correlated to the pK_i data using PLS. The model explained $R^2=0.91$ of the **y**-variance and showed high predictive ability for new protease-inhibitor combinations ($Q^2=0.82$) and for new inhibitors ($Q^2_{inh}=0.76$). In a further attempt to improve the model we added intraprotease cross-terms. This gave only a slight increase in the predictive ability ($Q^2=0.83$, $Q^2_{inh}=0.78$) indicating that a quite small fraction of all possible mutation pairs coexist and cooperate to reduce the compounds' inhibitory activities.

However, although the assessments of the two above models indicated that they performed statistically about equally well, we elected to use the more complex model incorporating the intraprotease cross-terms for predictions and interpretations. It is well-known that the presence of excessive irrelevant variables may introduce noise in PLS modeling, which is then manifested by a lowered Q^2. However, in our case possible negative effects of 'insignificant' intraprotease cross-terms (which actually point out noncooperative mutation pairs) were countervailed by relevant cross-terms representing epistatic effects between mutations. Hence, although the two models seemed to be equally predictive, we elected to analyze the one which allowed more detailed interpretations. Thus the model with intraprotease cross terms was the one used here in the following.

The goodness-of-fit and cross-validation results of the model are illustrated graphically in Figure 1. As seen from the figure, on the whole, over 95% of predictions fall in the area between the oblique gray lines representing an error of ± 1 log units. The good performance of the model is shown by its standard deviation of errors of prediction (SDEP) being 0.51. However, on a few occasions mispredictions by more than 1.5 logarithmic units occurred. These few mispredictions occurred for one of the compounds (Dmp-323; CID=72404), for which large discrepancies exist between the inhibition constant values reported in various sources.[20,21,24]

Although the model showed excellent predictive ability, it was based on interaction data for only 386 of all possible $21 \times 72 = 1512$ inhibitor-protease combinations. We therefore wanted to assess the relationships between the sparseness of the data matrix and the performance of the resulting proteochemometric models. Iterative modeling using twenty randomly selected fractions of sizes 25, 50, and 75% of all available data gave the following assessments of the external predictive ability: $P^2_{25\%}=0.62 \pm 0.06$, $P^2_{50\%}=0.74 \pm 0.04$, and $P^2_{75\%}=0.81 \pm 0.03$ (see Methods for details on calcula-

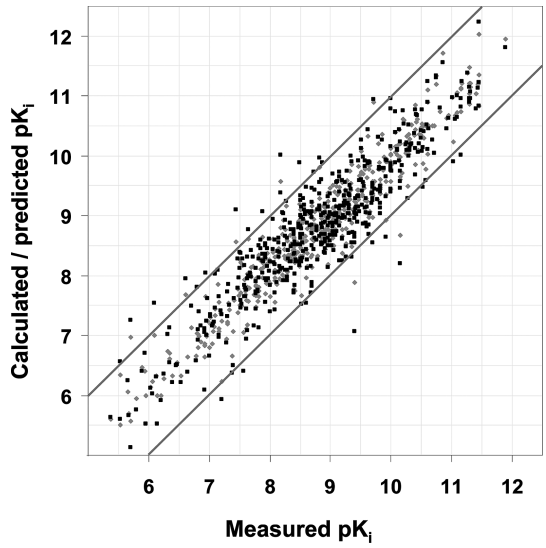

Figure 1. Correlation of calculated (gray diamonds) and predicted (black squares) pK_i versus measured pK_i values derived by the proteochemometric model for HIV proteases. Calculated values indicate the goodness of fit of the model; predicted values indicate the predictive ability as assessed by 7-fold cross-validation.

tions). These results show that even the models that were built on very sparse data matrices, comprising less than 100 interactions (i.e., on the average five protease mutant variants per compound and less than two compounds per protease variant), were still predictive. On the other hand, the P^2 increases when the size of the data set increases. The comparisons of the three P^2 values and their standard deviations thus emphasize the importance of well-populated data sets for predictive modeling and reliability of interpretations.

Cross-Resistance Conferring Mutations. The proteochemometric model was first used to identify mutations that decrease the activity of most of the inhibitors in the data set. Combinations of these mutations would bring to the worst scenario for the patient, i.e. a case when there may be no drug available that can effectively inhibit the virus. Accordingly, such mutations should be the ones to pay the most attention to in the design of new inhibitors.

Although we used a nonuniform approach to describe mutated sequence residues, a uniform way for interpretation is to compute the change in pK_i caused by each mutation. This was done by using the PLS regression equation. Thus, for each residue described by zz-scales we calculated the change in the average pK_i value over the whole compound series caused by a mutation by multiplying regression coefficients with the differences in descriptor values for the mutated and nonmutated protease according to the equation

$$\Delta \mathrm{p}K_i = \sum_{zz=1}^{3} ((coeff_{zz} + \sum_{p=1}^{P\times(P-1)/2} coeff_{p,zz} \times x_p) \times (x_{zz,mutated} - x_{zz,wild\text{-}type}))$$

where P indicates all other protease descriptors. In this way we assessed the change in inhibitory activity of the whole compounds series (i.e., average change), explained in the model by the three zz-scales and the intraprotease cross-terms formed there from.

Similarly, for each residue that showed variation of only two amino acids and was represented by a binary descriptor, the change in the pK_i value due to mutation was calculated as

$$\Delta \mathrm{p}K_i = (coeff_{bin} + \sum_{p=1}^{P\times(P-1)/2} coeff_{p,bin} \times x_p) \times (x_{bin,mutated} - x_{bin,wild\text{-}type})$$

The outcome of this analysis for 32 most frequent mutations is presented graphically in Figure 2. As seen from the figure seven sequence positions were identified with mutations giving a decrease in pK_i by more than 0.5 log units; among the most deleterious ones being the G48V, I50V, V82A, and I84V mutations. These aliphatic residues outline the active site in the protease, and here mutations are known to induce phenotypic resistance to the first seven inhibitors approved by FDA by modifying the geometry of the active site.[4] Peptide substrates seem to tolerate subtle conformational changes of the active site, while the more constrained rigid inhibitors may be incapable of adapting such geometric distortions and therefore lose their binding affinity.[46,47] For the active site-residue no. 82, substrates are known to tolerate a broad variation of chemical properties,[48] while our results in Figure 2 indicate that all three frequent mutations V82A, V82F, and V82T decrease inhibitors' activity.

Mutation D30N was previously known to diminish the susceptibility to nelfinavir.[4] However, our results show that this mutation causes a large decrease in K_i for the whole series of studied compounds. Calculating ΔpK_i for each particular compound in the series reveals that aside from nelfinavir the D30N mutation leads to a more than 10-fold decrease in K_i for several compounds, including the most recently approved inhibitor darunavir.

As can also be seen from Figure 2 many frequent mutations do not induce negative influence on inhibitors' activity. Some of these mutations, such as e.g. L10I, M36I, and I63P, are polymorphic, i.e. they reflect natural genetic variations.[48] On the other hand, several others occur in response to drug treatment and are associated with resistance (e.g., M46I, M46L, and I47V). Our analysis thus suggests that although these mutations are more prevalent in viruses

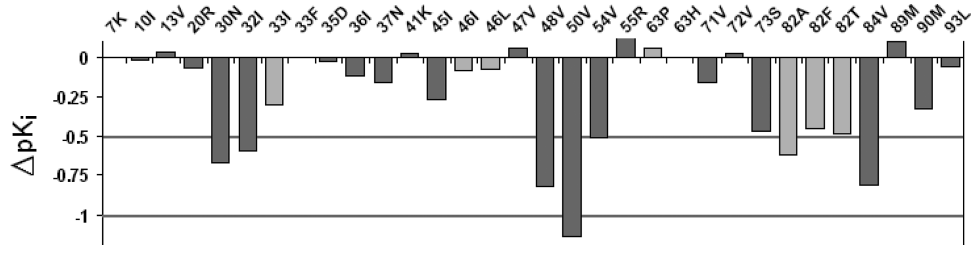

Figure 2. Calculated change in the average pK_i of the 21 inhibitors due to single mutations in the HIV protease.

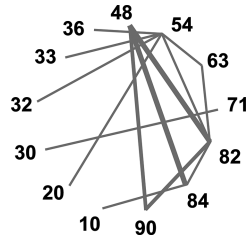

Figure 3. Schematic representation of the most significant intraprotease cross-terms. Each gray line represents a cross-term between the two sequence residues; the thickness of each line is proportional to the absolute value of the PLS regression coefficient.

from treated persons and might be beneficial for virus survival they do not directly interfere with inhibitor binding.

Cooperative Effects of Mutation Pairs. Although the introduction of intraprotease cross-terms resulted in only a slight improvement in modeling performance, several of these cross-terms obtained large positive or negative PLS coefficients. This indicates that mutation pairs with such significant cross-terms collaborate to reduce inhibitor activities. The most important cooperating mutations revealed by the modeling are presented schematically in Figure 3. As seen, the most negative PLS coefficients were obtained by the two cross-terms formed from the binary descriptor of residue 48 with the zz1- and zz3-scale descriptors of residue 82. The values of these two zz-scales are negative for Val, which is present at position 82 in the wild-type protease, and positive for Ala in the prevalent mutant V82A. The value of a binary descriptor is always positive if the described residue is mutated and negative (after centering) if not. Accordingly, the two cross-terms get high positive values for the double mutant G48V/V82A and negative values if only one of these residues is mutated. A negative PLS coefficient thus indicates that when the two mutations are present in combination the ΔpK_i becomes more negative than is expected from just summing up ΔpK_i for each of the two mutations separately (i.e., inhibitor affinities are reduced more than expected compared to the wild-type protease).

Large negative PLS coefficients were also given to the cross-terms between the binary descriptors of positions 48 and 84 (i.e., double mutant G48V/I84V) and the 48 and 90 positions (double mutant G48V/L90M). Further on, Figure 3 reveals that mutation of residue 54 (I54V) can amplify the effects of multiple other mutations.

The existence of cooperative coupling between distal protease sequence residues in the HIV protease was earlier described by Ohtaka et al.[49] E.g. these authors found that the double mutants V82A/I84V, M46I/I54V, and L10I/L90 M reduce the dissociation constant of indinavir by factors of 15, 1.5, and 3, respectively. However, when these were combined into a hexa-mutated protease a whole 200-fold decrease in affinity was observed. Similar results were obtained for each of the six inhibitors studied by Ohtaka et al. As illustrated in Figure 3, our present analysis shows that besides the existence of single cooperating pairs of sequence positions complex networks are present in the protease which can amplify the effects of individual mutations.

Inhibitor-Specific Mutations. In the further analysis we evaluated the influence of mutations on the activity of particular inhibitors. We first looked at the regression coefficients for protease-inhibitor cross-terms. As discussed above, these cross-terms explain why the same mutation may have a different effect depending on the properties of the inhibitor. To get an overview of the compound-distinguishing mutations, we summed the absolute values of coefficients for all protease-inhibitor cross-terms formed by each protease descriptor. The highest value for this cumulative measure was observed for the binary descriptor representing mutation G48V, followed by the second and third zz-scales characterizing mutations V82A/F/T (these z-scales describe size/shape and polarity of amino acids, respectively). High values were also found for the binary descriptors representing mutations L90M, V32I, I50V, and I54V (data not shown graphically). Thus, although these mutations were among the worst ones (*cf.* Figure 2) the analysis of cross-terms suggested that some compounds are affected less by them than others. The finding may give directions to the design of inhibitors with a broader spectrum of activity and was further illustrated by applying the model to predict pK_i for particular protease-inhibitor pairs. These predictions showed e.g. that the mutation G48V would give an over 10-fold decrease of K_i for saquinavir and Abbott77003 (CID=64999), while it would have quite a low influence on the activities of darunavir, amprenavir, TMC-126 (CID=200104), and VB-11328 (CID=462367). Similarly, predictions suggest that the V82F mutation is most detrimental for Dmp-323 (CID=72404), causing a more than 10-fold decrease of K_i, while it does not affect the activity of saquinavir or KNI-272 (CID=60927). Moreover, according to the model the L90 M mutation reduces the inhibitory activity of saquinavir by more than 0.5 log units, while it has essentially no effect on amprenavir, darunavir, and TMC126. The V32I mutation was found to be least detrimental for nelfinavir and saquinavir, while mutation I54V has the least influence on the activity of amprenavir, darunavir, and TMC-126.

Another noteworthy result is for the I50V mutation. This is the mutation which causes the overall most negative influence on pK_i for our data set compounds (c.f. Figure 2), and it is known as the key mutation conferring resistance to amprenavir as well as it reduces virus' susceptibility to ritonavir, lopinavir, nelfinavir, and saquinavir.[3] Our model predicts that for the most recently approved inhibitor darunavir loss of activity due to this amino acid substitution would be as large as $\Delta pK_i = -1.4$. This is in a quite good agreement with newly reported experimentally measured activities for darunavir. Thus, a recent study by Wang et al. reported a 3.9- fold activity decrease for the I50V mutation compared to wild-type protease,[50] while Liu *et al.*[51] found that the activity of darunavir dropped by more than 30-fold from 0.59 nM for the wild-type protease to 18 nM for this mutant. Although some discrepancies certainly do exist between the data collected from several sources, it thus seems that the general trends captured by our model are reliable.

Contributions of Molecular Properties of Chemical Compounds to the Inhibition of Multimutated Proteases. The analysis may be scrutinized further by simultaneously analyzing coefficients for intraprotease and protease-inhibitor cross-terms to characterize co-operative effects of amino acid mutations in various combinations on the activities of particular inhibitors. Although a complete analysis is beyond any practical possibilities in a written report like this, bearing in mind the immense number of possible mutation combinations, we will demonstrate the feasibility of using the model to interpret interactions of inhibitors with proteases bearing

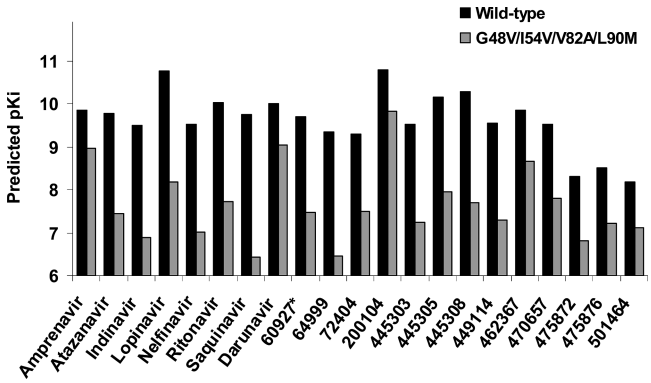

Figure 4. Predicted inhibitory activity (pK_i) against wild-type HIV protease and its mutant variant G48V/I54V/V82A/L90M. For nonapproved inhibitors shown are PubChem compound accession identifier (CID) numbers.

essentially any clinically known or hypothetical mutation pattern by using a case example.

Using the PLS regression coefficients of the model the contribution of a compound property represented by a molecular descriptor l for the inhibition of a protease A can be calculated according to the equation

$$\Delta \mathrm{p}K_i(l,A) = coeff_l + \sum_{p=1}^{P} (coeff_{l,p} x_{p,A})$$

where $\Delta \mathrm{p}K_i(l,A)$ is change of inhibition constant when the value of descriptor l increases by one standard deviation, $coeff_l$ and $coeff_{l,p}$ are regression coefficients for descriptor l and its cross-terms with protease descriptors $p = 1$ to P, and $x_{p,A}$ are the actual values of the descriptors for protease A.

In our example we will compare descriptor contributions for the inhibition of the wild-type protease and the protease combining the most common resistance conferring mutations, namely G48V/I54V/V82A/L90M. This combination of amino acid substitutions is known to lead to more than a 100-fold reduction of virus susceptibility to saquinavir, 20- to 50-fold reductions of the susceptibilities to atazanavir, indinavir, lopinavir, nelfinavir, and ritonavir, while there is an only a 2-fold reduction of susceptibility to amprenavir (data for darunavir are not available).[3] The predicted pK_i values for the 21 compounds in our data set suggest that along with amprenavir, several other compounds, namely TMC-126 (CID=200104), darunavir, and VB-11328 (CID=462367), should also retain high activities for the tetramutated protease (see Figure 4).

In Figure 5 is further shown graphically the calculated contribution values for the 84 molecular descriptors of the protease inhibitors used in the model for the inhibition of wild-type and tetramutated protease. As can be seen molecular weight correlates subtly negatively to the activity against both the wild-type and mutated protease. Descriptors characterizing molecular flexibility (such as number of circuits and rings, rotatable and double bonds) do not correlate with the activity of our data set compounds. However, an important correlation is seen for molecular shape, represented by descriptors such as 3D-Balaban index (J3D), spherosity (SPH), molecular eccentricity (MEcc), and 3D Petitjean shape index (PJI3). Figure 5 also reveals that large maximum negative charge and relative negative charge of the molecule (represented by qneg and RNCG descriptors, respectively) are favorable for the inhibition of both the wild-type and mutated protease. Moreover, as indicated by the negative values for the nR07 descriptor, lower activity was possessed by compounds containing a seven-membered ring, which in our data set was present in cyclic urea and cyclic sulfamide derivatives. Comparing these two scaffolds, the presence of a sulfonyl group was found to be more favorable than carbonyl, as signified by positive $\Delta \mathrm{p}K_i$ values for the nSO2 descriptor.

High contribution values to the inhibition of both proteases was further given to the fragment O-59, which is part of the tetrahydrofuran moiety in amprenavir and VB-11328 and bis-tetrahydrofuran in darunavir and TMC-126 (CID=200104). The advantage of the latter moiety for high inhibitory activity was also confirmed by a positive influence on $\Delta \mathrm{p}K_i$ values for nR08 descriptor. Very positively assessed is also the fragment H-053, which represents hydrogen attached to C(sp3) with formal oxidation number 0 (i.e., not being connected to electronegative atoms) having two electronegative atoms attached to the carbon next to it; such a fragment is, for example, present in the central part of the ritonavir and lopinavir structures.

By contrast to the above findings, several other descriptors obtained substantially different $\Delta \mathrm{p}K_i$ values for the two protease variants. A sulfonamide group (nSO2N), which is present in darunavir and amprenavir, was found to be particularly good for inhibition of the tetramutated protease. The descriptor H-051, which represents the number of hydrogens attached to alpha-C atoms in the molecule, correlates on the other hand with a lower $\Delta \mathrm{p}K_i$ value for the mutated than for the wild-type protease. Differences are also seen for the MAXDN descriptor, which relates to nucleophilicity,[52] and the qpos descriptor which represents the maximum positive charge of the molecule.

The overall patterns of the descriptor contributions for the inhibition of the wild-type and mutated protease are nevertheless quite similar; the squared correlation coefficient between the two series of values being 0.76. In fact this is quite expected since the development of resistance is a complex process which one would be naive to think that it could be attributed to one or few unfavorable molecular properties of the inhibitors. Still the accumulated differences from all descriptors allowed the model to explain the more than a 1000-fold activity decrease for some compounds compared to less than a 10-fold decrease for some others by the combined G48V/I54V/V82A/L90M mutations (see the predictions in Figure 4). Thus, the whole set of descriptors captured the molecular properties that determine compound interactions with the mutated virus variants. Such a joint explanation of inhibitor-protease interactions by multiple descriptors makes the model more robust (compared to a model where only few descriptors would appear important) and would increase the 'chemical space' where accurate predictions for novel compounds could be performed. In other words the model should be useful to predict the activities of modified or completely novel *in silico* designed structures.

In this context it should also be mentioned that the resolution of proteochemometric models generally depends on the number of interacting entities and the diversity of their interaction profiles in the data set. We should thus anticipate further improvements of the model when data for

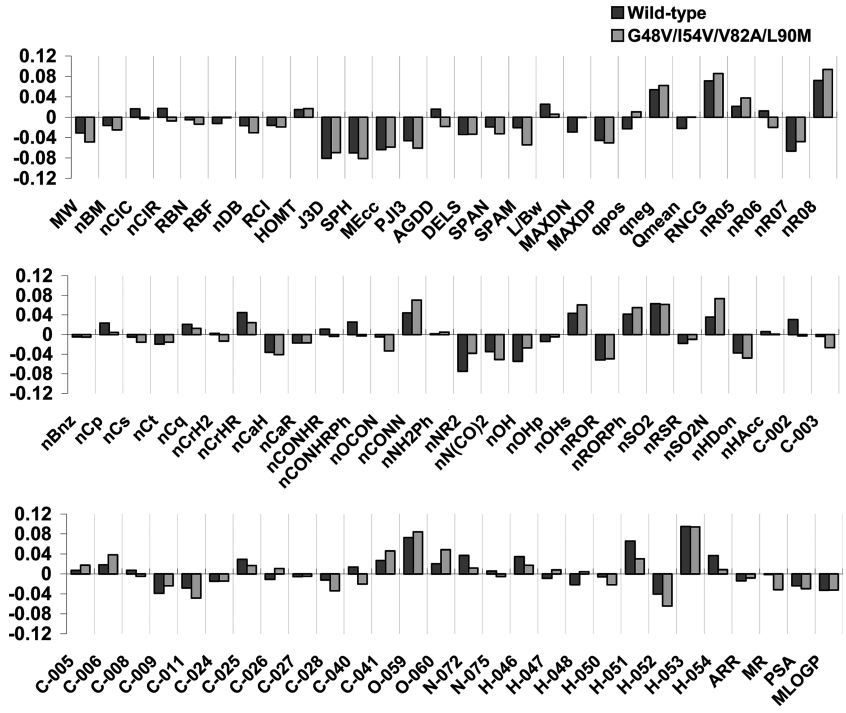

Figure 5. Contributions of 84 molecular properties for the inhibition of the wild-type HIV protease and its mutant variant G48V/I54V/V82A/L90M. Contributions represent changes in calculated inhibitory activity (ΔpK_i) when inhibitor property is modified so that the descriptor value increases by one standard deviation. Abbreviations correspond to descriptors as follows: MW: molecular weight; nBM: number of multiple bonds; nCIC: number of rings; nCIR: number of circuits; RBN: number of rotatable bonds; RBF: rotatable bond fraction; nDB: number of double bonds; RCI: Jug RC index; HOMT: HOMA total aromaticity index; J3D: 3D-Balaban index; SPH: spherosity; MEcc: molecular eccentricity; PJI3: 3D Petitjean shape index; AGDD: average geometric distance degree; DELS: molecular electrotopological variation; SPAN: span R; SPAM: average span R; L/Bw: length-to-breadth ratio by WHIM; MAXDN: maximal electrotopological negative variation; MAXDP: maximal electrotopological positive variation; qpos: maximum positive charge; qneg: maximum negative charge; Qmean: mean absolute charge (charge polarization); RNCG: relative negative charge; nR05: number of 5-membered rings; nR06: number of 6-membered rings; nR07: number of 7-membered rings; nR08: number of 8-membered rings; nBnz: number of benzene-like rings; nCp: number of terminal primary C(sp3); nCs: number of total secondary C(sp3); nCt: number of total tertiary C(sp3); nCq: number of total quaternary C(sp3); nCrH2: number of ring secondary C(sp3); nCrHR: number of ring tertiary C(sp3); nCaH: number of unsubstituted aromatic C(Sp2); nCaR: number of substituted aromatic C(Sp2); nCONHR: number of secondary amides (aliphatic); nCONHRPh: number of secondary amides (aromatic); nOCON: number of carbamates (aliphatic); nCONN: number of urea derivatives; nNH2Ph: number of primary amines (aromatic); nNR2: number of tertiary amines (aliphatic); nN(CO)2: number of imides; nOH: number of total hydroxyl groups; nOHp: number of primary alcohols (aliphatic); nOHs: number of secondary alcohols (aliphatic); nROR: number of ethers (aliphatic); nRORPh: number of ethers (aromatic); nSO2: number of sulfones; nRSR: number of solfures; nSO2N: number of sulfonamides; nHDon: number of donor atoms for H-bonds; nHAcc: number of acceptor atoms for H-bonds; C-002: fragment CH2R2; C-003: fragment CHR3; C-005: fragment CH3X; C-006: fragment CH2RX; C-008: fragment CHR2X; C-009: fragment CHRX2; C-011: fragment CR3X; C-024: fragment R--CH--R; C-025: fragment R--CR--R; C-026: fragment R--CX--R; C-027: fragment R--CH--X; C-028: fragment R--CR--X; C-040: fragment R-C(=X)-X/R-C#X/X=C=X; C-041: fragment X-C(=X)-X; O-059: fragment Al−O−Al; O-060: fragment Al−O−Ar/Ar−O−Ar/R..O..R/R−O-C=X; N-072: fragment RCO-N</>N-X=X; N-075: fragment R--N--R/R--N--X; H-046: H attached to C^0(sp3)no X attached to next C; H-047: H attached to C^1(sp3)/ C^0(sp2); H-048: H attached to C^2(sp3)/C^1(sp2)/C^0(sp); H-050: H attached to heteroatom; H-051: H attached to alpha-C; H-052: H attached to C^0(sp3) with 1X attached to next C; H-053: H attached to C^0(sp3) with 2X attached to next C; H-054: H attached to C^0(sp3) with 3X attached to next C; ARR: aromatic ratio; MR: Ghose-Crippen molar refractivity; PSA: fragment-based polar surface area; MLOGP: Moriguchi octanol−water partition coefficient. Notation of atom and bond types in fragment descriptors corresponds to Viswanadhan et al.[55]

more compounds showing more diverse inhibition patterns become available. At present activity data are collected *en masse* only for the wild-type protease (see e.g. Binding DB and NIAID ChemDB databases[53,54]). However, as there is a continuous clinical need for retardants against viral strains that have developed resistance to older agents more and more data for protease mutants will become amenable to allow extended proteochemometric modeling in the near future, something that we predict will greatly increase its ability to direct the development of new improved anti-HIV agents.

CONCLUSIONS

We have performed simultaneous analysis of the inhibitory activity of a series of 21 compounds with 72 HIV protease variants using publicly available data from more than ten years of HIV research. Our proteochemometric model explained 92% of the variation in pK_i of inhibitor-protease pairs, and it was highly predictive. The model revealed the mutations that have a large negative influence on the pK_i of the majority of the studied compounds, namely D30N, V32I, G48V, I50V, I54V, V82A, I84V, and L90M. At the same time, several mutations that are known to reduce HIV drug susceptibility did not appear as being important in our model, suggesting that their role is rather to exert compensatory effects on substrate cleavage.

The model also revealed cooperative effects of certain pairs of mutations, and the analysis could even be extended beyond that to cover the impact of larger sets of amino acid mutations in multimutated proteases. In addition to the general analysis of the impact of mutations to the overall activity of the whole

series of compounds, our model reveals the influence of mutations on particular inhibitors and predicts pK_i values for particular inhibitor-protease combinations. This showed, for example, that although mutations of residues such as 48, 84, and 90 were generally very detrimental to the compounds' activities, several inhibitors were identified that for each of these mutations did not suffer a substantial loss of activity. Finally, the model also gave insights into the molecular properties of the compounds that are important for inhibition of wild-type and mutated proteases.

To sum up, proteochemometric modeling proves to be well suited for the evaluation and prediction of inhibitor interactions with mutated HIV protease variants. Modeling results can potentially be used for selection of compounds with improved HIV retarding properties showing improved ability to surmount resistance. Our modeling approach is general and can incorporate an unlimited amount of data for mutated targets and organic compounds which would lead to improved model performance. Moreover, our approach is completely general and can be adapted to the analysis of drug interactions with virtually any heterogeneous target.

ACKNOWLEDGMENT

This research was supported by the Swedish VR (04X-05957), Stiftelsen Läkare mot AIDS Forskningsfond, and Åke Wibergs Stiftelse.

Supporting Information Available: Data set used for construction of the proteochemometric model. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Saksena, N. K.; Potter, S. J. Reservoirs of HIV-1 in vivo: implications for antiretroviral therapy. *AIDS Rev.* **2003**, *5*, 3–18.

(2) Nijhuis, M.; van Maarseveen, N. M.; Lastere, S.; Schipper, P.; Coakley, E.; Glass, B.; Rovenskav, M.; de Jongv, D.; Chappey, C.; Goedegebuure, I. W.; Heilek-Snyder, G.; Dulude, D.; Cammack, N.; Brakier-Gingras, L.; Konvalinka, J.; Parkin, N.; Kräusslich, H. G.; Brun-Vezinet, F.; Boucher, C. A. A novel substrate-based HIV-1 protease inhibitor drug resistance mechanism. *PLoS Med.* **2007**, *4*, e36.

(3) Rhee, S. Y.; Gonzales, M. J.; Kantor, R.; Betts, B. J.; Ravela, J.; Shafer, R. W. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* **2003**, *31*, 298–303.

(4) Lapins, M.; Eklund, M.; Spjuth, O.; Prusis, P.; Wikberg, J. E. Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinf.* **2008**, *5*, Article 181. http://www.biomedcentral.com/1471-2105/9/181 (accessed Feb 28, 2009).

(5) 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.

(6) Kiralj, R.; Ferreira, M. M. A priori molecular descriptors in QSAR: a case of HIV-1 protease inhibitors. I. The chemometric approach. *J. Mol. Graphics Modell.* **2003**, *21*, 435–448.

(7) Kurup, A.; Mekapativ, S. B.; Garg, R.; Hansch, C. HIV-1 protease inhibitors: a comparative QSAR analysis. *Curr. Med. Chem.* **2003**, *210*, 1679–1688.

(8) Sirois, S.; Tsoukas, C. M.; Chou, K. C.; Wei, D.; Boucher, C.; Hatzakis, G. E. Selection of molecular descriptors with artificial intelligence for the understanding of HIV-1 protease peptidomimetic inhibitors-activity. *Med. Chem.* **2005**, *1*, 173–184.

(9) Debnath, A. K. Application of 3D-QSAR techniques in anti-HIV-1 drug design--an overview. *Curr. Pharm. Des.* **2005**, *11*, 3091–3110.

(10) Avram, S.; Bologa, C.; Flonta, M. L. Quantitative structure-activity relationship by CoMFA for cyclic urea and nonpeptide-cyclic cyanoguanidine derivatives on wild type and mutant HIV-1 protease. *J. Mol. Model.* **2005**, *11*, 105–115.

(11) Almerico, A. M.; Tutone, M.; Lauria, A.; Diana, P.; Barraja, P.; Montalbano, A.; Cirrincione, G.; Dattolo, G. A multivariate analysis of HIV-1 protease inhibitors and resistance induced by mutation. *J. Chem. Inf. Model.* **2006**, *46*, 168–179.

(12) Wikberg, J. E.; Lapinsh, M. ; Prusis, P. Proteochemometrics: A tool for modelling the molecular interaction space. In *Chemogenomics in Drug Discovery - A Medicinal Chemistry Perspective*; Kubinyi, H., Müller, G., Eds.; Wiley-VCH: Weinheim, 2004; pp 289−309.

(13) Lapinsh, M.; Prusis, P.; Petrovska, R.; Uhlén, S.; Mutule, I.; Veiksina, S.; Wikberg, J. E: Proteochemometric modeling reveals the interaction site for Trp9 modified a-MSH peptides in melanocortin receptors. *Proteins* **2007**, *67*, 653–660.

(14) Lapinsh, M.; Prusis, P.; Uhlén, S.; Wikberg, J. E. Improved approach for proteochemometrics modeling: application to organic compound - amine G protein-coupled receptor interactions. *Bioinformatics* **2005**, *21*, 4289–4296.

(15) Lapinsh, M.; Veiksina, S.; Uhlén, S.; Petrovska, R.; Mutule, I.; Mutulis, F.; Yahorava, S.; Prusis, P.; Wikberg, J. E. Proteochemometric mapping of the interaction of organic compounds with melanocortin receptor subtypes. *Mol. Pharm.* **2005**, *67*, 50–59.

(16) Vacca, J. P.; Dorsey, B. D.; Schleif, W. A.; Levin, R. B.; McDaniel, S. L.; Darke, P. L.; Zugay, J.; Quintero, J. C.; Blahy, O. M.; Roth, E.; et al. L-735,524: an orally bioavailable human immunodeficiency virus type 1 protease inhibitor. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *26*, 4096–4100.

(17) Kaplan, A. H.; Michael, S. F.; Wehbie, R. S.; Knigge, M. F.; Paul, D. A.; Everitt, L.; Kempf, D. J.; Norbeck, D. W.; Erickson, J. W.; Swanstrom, R. Selection of multiple human immunodeficiency virus type 1 variants that encode viral proteases with decreased sensitivity to an inhibitor of the viral protease. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 5597–5601.

(18) Ho, D. D.; Toyoshima, T.; Mo, H.; Kempf, D. J.; Norbeck, D.; Chen, C. M.; Wideburg, N. E.; Burt, S. K.; Erickson, J. W.; Singh, M. K. Characterization of human immunodeficiency virus type 1 variants with increased resistance to a C2-symmetric protease inhibitor. *J. Virol.* **1994**, *68*, 2016–2020.

(19) Partaledis, J. A.; Yamaguchi, K.; Tisdale, M.; Blair, E. E.; Falcione, C.; Maschera, B.; Myers, R. E.; Pazhanisamy, S.; Futer, O.; Cullinan, A. B.; et al. In vitro selection and characterization of human immunodeficiency virus type 1 (HIV-1) isolates with reduced sensitivity to hydroxyethylamino sulfonamide inhibitors of HIV-1 aspartyl protease. *J. Virol.* **1995**, *69*, 5228–5235.

(20) Nillroth, U.; Vrang, L.; Markgren, P. O.; Hultén, J.; Hallberg, A.; Danielson, U. H. Human immunodeficiency virus type 1 proteinase resistance to symmetric cyclic urea inhibitor analogs. *Antimicrob. Agents Chemother.* **1997**, *41*, 2383–2388.

(21) Wilson, S. I.; Phylip, L. H.; Mills, J. S.; Gulnik, S. V.; Erickson, J. W.; Dunn, B. M.; Kay, J. Escape mutants of HIV-1 proteinase: enzymic efficiency and susceptibility to inhibition. *Biochim. Biophys. Acta* **1997**, *1339*, 113–125.

(22) Sham, H. L.; Kempf, D. J.; Molla, A.; Marsh, K. C.; Kumar, G. N.; Chen, C. M.; Kati, W.; Stewart, K.; Lal, R.; Hsu, A.; Betebenner, D.; Korneyeva, M.; Vasavanonda, S.; McDonald, E.; Saldivar, A.; Wideburg, N.; Chen, X.; Niu, P.; Park, C.; Jayanti, V.; Grabowski, B.; Granneman, G. R.; Sun, E.; Japour, A. J.; Leonard, J. M.; Plattner, J. J.; Norbeck, D. W. ABT-378, a highly potent inhibitor of the human immunodeficiency virus protease. *Antimicrob. Agents Chemother.* **1998**, *42*, 3218–3224.

(23) Gulnik, S. V.; Suvorov, L. I.; Liu, B.; Yu, B.; Anderson, B.; Mitsuya, H.; Erickson, J. W. Kinetic characterization and cross-resistance patterns of HIV-1 protease mutants selected under drug pressure. *Biochemistry* **1995**, *34*, 9282–8287.

(24) Ala, P. J.; Huston, E. E.; Klabe, R. M.; McCabe, D. D.; Duke, J. L.; Rizzo, C. J.; Korant, B. D.; DeLoskey, R. J.; Lam, P. Y.; Hodge, C. N.; Chang, C. H. Molecular basis of HIV-1 protease drug resistance: structural analysis of mutant proteases complexed with cyclic urea inhibitors. *Biochemistry* **1997**, *36*, 1573–1580.

(25) Pivazyan, A. D.; Matteson, D. S.; Fabry-Asztalos, L.; Singh, R. P.; Lin, P. F.; Blair, W.; Guo, K.; Robinson, B.; Prusoff, W. H. Inhibition of HIV-1 protease by a boron-modified polypeptide. *Biochem. Pharmacol.* **2000**, *60*, 927–936.

(26) Dorsey, B. D.; McDonough, C.; McDaniel, S. L.; Levin, R. B.; Newton, C. L.; Hoffman, J. M.; Darke, PvL.; Zugay-Murphy, J. A.; Emini, E. A.; Schleif, W. A.; Olsen, D. B.; Stahlhut, M. W.; Rutkowski, C. A.; Kuo, L. C.; Lin, J. H.; Chen, I. W.; Michelson, S. R.; Holloway, M. K.; Huff, J. R.; Vacca, J. P. Identification of MK-944a: a second clinical candidate from the hydroxylaminepentanamide isostere series of HIV protease inhibitors. *J. Med. Chem.* **2000**, *43*, 3386–3399.

(27) Schock, H. B.; Garsky, V. M.; Kuo, L. C. Mutational anatomy of an HIV-1 protease variant conferring cross-resistance to protease inhibitors in clinical trials. Compensatory modulations of binding and activity. *J. Biol. Chem.* **1996**, *271*, 31957–31963.

(28) Kovalevsky, A. Y.; Tie, Y.; Liu, F.; Boross, P. I.; Wang, Y. F.; Leshchenko, S.; Ghosh, A. K.; Harrison, R. W.; Weber, I. T.

Effectiveness of nonpeptide clinical inhibitor TMC-114 on HIV-1 protease with highly drug resistant mutations D30N, I50V, and L90M. *J. Med. Chem.* **2006**, *49*, 1379–1387.
(29) Weber, J.; Mesters, J. R.; Lepsík, M.; Prejdová, J.; Svec, M.; Sponarová, J.; Mlcochová, P.; Skalická, K.; Strísovský, K.; Uhlíková, T.; Soucek, M.; Machala, L.; Stanková, M.; Vondrásek, J.; Klimkait, T.; Kraeusslich, H. G.; Hilgenfeld, R.; Konvalinka, J. Unusual binding mode of an HIV-1 protease inhibitor explains its potency against multi-drug-resistant virus strains. *J. Mol. Biol.* **2002**, *324*, 739–754.
(30) Ali, A.; Reddy, G. S.; Cao, H.; Anjum, S. G.; Nalam, M. N.; Schiffer, C. A.; Rana, T. M. Discovery of HIV-1 protease inhibitors with picomolar affinities incorporating N-aryl-oxazolidinone-5-carboxamides as novel P2 ligands. *J. Med. Chem.* **2006**, *49*, 7342–7356.
(31) Ahlsén, G.; Hultén, J.; Shuman, C. F.; Poliakov, A.; Lindgren, M. T.; Alterman, M.; Samuelsson, B.; Hallberg, A.; Danielson, U. H. Resistance profiles of cyclic and linear inhibitors of HIV-1 protease. *Antivir. Chem. Chemother.* **2002**, *13*, 27–37.
(32) Clemente, J. C.; Moose, R. E.; Hemrajani, R.; Whitford, L. R.; Govindasamy, L.; Reutzel, R.; McKenna, R.; Agbandje-McKenna, M.; Goodenow, M. M.; Dunn, B. M. Comparing the accumulation of active- and nonactive-site mutations in the HIV-1 protease. *Biochemistry* **2004**, *43*, 12141–12151.
(33) Clemente, J. C.; Coman, R. M.; Thiaville, M. M.; Janka, L. K.; Jeung, J. A.; Nukoolkarn, S.; Govindasamy, L.; Agbandje-McKenna, M.; McKenna, R.; Leelamanit, W.; Goodenow, M. M.; Dunn, B. M. Analysis of HIV-1 CRF_01 A/E protease inhibitor resistance: structural determinants for maintaining sensitivity and developing resistance to atazanavir. *Biochemistry* **2006**, *45*, 5468–5477.
(34) Clemente, J. C.; Hemrajani, R.; Blum, L. E.; Goodenow, M. M.; Dunn, B. M. Secondary mutations M36I and A71V in the human immunodeficiency virus type 1 protease can provide an advantage for the emergence of the primary mutation D30N. *Biochemistry* **2003**, *42*, 15029–15035.
(35) Ghosh, A. K.; Ramu Sridhar, P.; Kumaragurubaran, N.; Koh, Y.; Weber, I. T.; Mitsuya, H. Bis-tetrahydrofuran: a privileged ligand for darunavir and a new generation of hiv protease inhibitors that combat drug resistance. *ChemMedChem.* **2006**, *1*, 939–950.
(36) Reddy, G. S.; Ali, A.; Nalam, M. N.; Anjum, S. G.; Cao, H.; Nathans, R. S.; Schiffer, C. A.; Rana, T. M. Design and synthesis of HIV-1 protease inhibitors incorporating oxazolidinones as P2/P2' ligands in pseudosymmetric dipeptide isosteres. *J. Med. Chem.* **2007**, *50*, 4316–4328.
(37) 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.
(38) Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab.* **1987**, *2*, 37–52.
(39) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, 2000.
(40) Geladi, P.; Kowalski, B. R. Partial least-squares regression: a tutorial. *Anal. Chim. Acta* **1986**, *185*, 1–17.
(41) Wold, S.; Trygg, J.; Berglund, A.; Antti, H. Some recent developments in PLS modeling. *Chemom. Intell. Lab.* **2001**, *58*, 131–150.
(42) Trygg, J.; Wold, S. Orthogonal projections to latent structures (O-PLS). *J. Chemom.* **2002**, *16*, 119–128.
(43) Spjuth, O.; Helmus, T.; Willighagen, E. L.; Kuhn, S.; Eklund, M.; Wagener, J.; Murray-Rust, P.; Steinbeck, C.; Wikberg, J. E. Bioclipse: An open source workbench for chemo- and bioinformatics. *BMC Bioinf.* **2007**, *8:59*.
(44) Genetta Soft. http://www.genettasoft.com (accessed Feb 28, 2009).
(45) Wold, S. PLS for multivariate linear modeling. In *Chemometric Methods in Molecular Design*; van de Waterbeemd, H., Ed.; VCH: Weinheim, 1995; pp 195−218.
(46) Luque, I.; Todd, M. J.; Gómez, J.; Semo, N.; Freire, E. Molecular basis of resistance to HIV-1 protease inhibition: a plausible hypothesis. *Biochemistry* **1998**, *37*, 5791–5797.
(47) Ohtaka, H.; Freire, E. Adaptive inhibitors of the HIV-1 protease. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 193–208.
(48) Rhee, S.; Fessel, W.; Zolopa, A.; Hurley, L.; Liu, T.; Taylor, J.; Nguyen, D.; Slome, S.; Klein, D.; Horberg, M.; et al. 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.
(49) Ohtaka, H.; Schön, A.; Freire, E. Multidrug resistance to HIV-1 protease inhibition requires cooperative coupling between distal mutations. *Biochemistry* **2003**, *42*, 13659–13666.
(50) Wang, Y. F.; Tie, Y.; Boross, P. I.; Tozser, J.; Ghosh, A. K.; Harrison, R. W.; Weber, I. T. Potent new antiviral compound shows similar inhibition and structural interactions with drug resistant mutants and wild type HIV-1 protease. *J. Med. Chem.* **2007**, *50*, 4509–4515.
(51) Liu, F.; Kovalevsky, A. Y.; Tie, Y.; Ghosh, A. K.; Harrison, R. W.; Weber, I. T. Effect of flap mutations on structure of HIV-1 protease and inhibition by saquinavir and darunavir. *J. Mol. Biol.* **2008**, *381*, 102–115.
(52) Gramatica, P.; Corradi, M.; Consonni, V. Modelling and prediction of soil sorption coefficients of non-ionic organic pesticides by molecular descriptors. *Chemosphere* **2000**, *41*, 763–777.
(53) Liu, T.; Lin, Y.; Wen, X; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35*, D198−201.
(54) HIV/OI Chemical Biological Database. http://chemdb.niaid.nih.gov (accessed Feb 28, 2009).
(55) Viswanadhan, V. N.; Ghose, A. K.; Revankar, G. R.; Robins, R. K. Atomic physicochemical parameters for three dimensional structure directed quantitative structure-activity relationships. 4. Additional parameters for hydrophobic and dispersive interactions and their application for an automated superposition of certain naturally occurring nucleoside antibiotics. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 163–172.

CI800453K