A Look Inside HIV Resistance through Retroviral Protease Interaction Maps

Aleksejs Kontijevskis1,2, Peteris Prusis1, Ramona Petrovska1, Sviatlana Yahorava1, Felikss Mutulis1, Ilze Mutule1, Jan Komorowski2, Jarl E. S. Wikberg1*  

Abstract

Retroviruses affect a large number of species, from fish and birds to mammals and humans, with global socioeconomic negative impacts. Here the authors report and experimentally validate a novel approach for the analysis of the molecular networks that are involved in the recognition of substrates by retroviral proteases. Using multivariate analysis of the sequence-based physicochemical descriptions of 61 retroviral proteases comprising wild-type proteases, natural mutants, and drug-resistant forms of proteases from nine different viral species in relation to their ability to cleave 299 substrates, the authors mapped the physicochemical properties and cross-dependencies of the amino acids of the proteases and their substrates, which revealed a complex molecular interaction network of substrate recognition and cleavage. The approach allowed a detailed analysis of the molecular–chemical mechanisms involved in substrate cleavage by retroviral proteases.

Introduction

Retroviruses are associated with a broad range of diseases that include tumors, immunodeficiency syndromes, and neurological disorders [1]. They affect a large number of species, from fish and birds to mammals and humans, with global socioeconomic negative impacts [1]. Each year the HIV pandemic causes more than 3 million deaths despite advances in the development of anti-HIV therapies [2]. The seemingly endless capability of retroviruses to escape antiviral drugs undermines treatment strategies and prompts the need for new broad-spectrum therapeutic agents [3].

Retroviral proteases process viral precursor polyproteins into structurally and functionally mature proteins that combine into infectious viral forms. As such, these proteins are key targets for the design of therapeutic inhibitors [4,5]. To date, the majority of protease inhibitors for treatment of HIV have been peptide mimetics, and most of them were specifically designed against only one of the HIV-1 proteases, namely the HXB2 (“wild-type”) HIV-1 protease [6,7]. Unfortunately, this strategy has led to failures to retard the replication of strains bearing drug-resistant protease mutations [3,8].

Although efficiently hydrolysable protease substrates have served as excellent templates for peptide-mimetic inhibitor design, it is difficult to predict which combination of amino acids will make the best substrate over multiple proteases [6]. Analysis of protease mutations associated with drug resistance is also confounded by the existence of many viral subtypes carrying naturally occurring polymorphisms [9]. The genomic differences among HIV-1 proteases can be as high as 30% and range from 10%–70% within the retroviral protease class [3]. Mutations contributing to viral resistance to antiviral drugs in one particular HIV subtype are found frequently in equivalent positions in the genes of other HIV subtypes or other retroviral proteases [9–14]. Still, the roles of specific mutations are only partly understood [5].

Here we report the development and experimental validation of a novel strategy for the molecular analysis of retroviral proteases. We hypothesized that merging essentially all available knowledge of retroviral proteases and their interactions with their substrates into a unified model would provide broad insight into the function of these enzymes and facilitate the analysis of retroviral drug resistance mechanisms. The modeling that we here report is based on the multivariate analysis of sequence position–physicochemical properties of the amino acids of 61 retroviral proteases from nine viral species and reveals a complex network of physicochemical interactions involved in protease recognition and cleavage of substrates. The approach provides novel insights into the molecular mechanisms involved in substrate cleavage by retroviral proteases in general as well as in relation to drug resistance.

Results

Substrate CRM for Retroviral Proteases

The model was based on an extensive survey of publicly available data from multiple retroviral proteases and their
Author Summary

Retroviruses are associated with a broad range of diseases that include tumor formation, neurological disorders, and immunodeficiency syndromes, including those of HIV. The extraordinary mutational plasticity of HIV-1 causes the rapid appearance of highly diverse quasi-species in a very short time, leading to severe problems with drug resistance. We here present and validate experimentally a novel approach for the analysis of the molecular interaction networks involved in the recognition process of substrates by natural and drug-resistant retroviral proteases. By combining a large number of wild-type and mutant retroviral proteases from nine different viral species, and their interactions with a large number of substrates, we have created a unified model incorporating all the proteases’ mutational space. Our results reveal that a complex physicochemical interaction network is involved in substrate recognition and cleavage by aspartate proteases and unravel detailed molecular mechanisms involved in drug resistance. These findings provide novel implications for understanding important features of HIV resistance and raise the possibility of developing completely novel strategies for the design of protease inhibitors that will remain effective over time despite rapid viral evolution.

Table 1. Creation of a Substratethe CRM for Retroviral Proteases

| Model | R²  | Q²  | RMSEE | Descriptors Included in the Model | Observations—Coefficient Ratio |
|-------|-----|-----|-------|-----------------------------------|-------------------------------|
| 1     | 0.38| 0.29| 0.79  | B, C                             | 1:0.6                         |
| 2     | 0.48| 0.36| 0.72  | B, C, B × C                      | 1:23.0                        |
| 3     | 0.60| 0.40| 0.63  | B, C, C × C                      | 1:1.6                         |
| 4     | 0.72| 0.51| 0.53  | B, C, B × C, C × C              | 1:24.0                        |
| 5     | 0.40| 0.33| 0.77  | A, B, C                         | 1:0.6                         |
| STM   | 0.56| 0.36| 0.64  | C, C × C                        | 1:3.9                         |
| CRM   | 0.77| 0.52| 0.49  | A, B, C, A × A, A × B, A × C, B × C, C × C | 1:28.9 |

The CRM was developed by inclusion of different descriptor blocks, A–C, until the best model was obtained. The descriptor blocks were as follows: A, assay constituents descriptor block; B, protease descriptor block; C, substrate descriptor block. A × A, A × B, A × C, B × C, and C × C represent the cross-term blocks formed from respective descriptor block. As seen in the table, the CRM, including all descriptor blocks, (except B × B, which was excluded due to its size not to induce overfitting), outperformed models 1–5 and STM (i.e., the latter is the STM representing a model for the HXB2 HIV-1 protease only). Ratio “Observations—Coefficient Ratio” denotes the fraction of the number of observations included in the model versus the number of descriptors used for model construction.

doi:10.1371/journal.pcbi.0030048.t001

External Validation of CRM

To validate the model further we examined its capacity to predict the activity of naturally occurring and artificially mutated retroviral proteases externally. This was afforded by excluding all data for eight retroviral strains one at a time in their entirety, and then predicting the excluded data using models constructed from the remaining data (see Materials and Methods for details). This analysis showed that the models could accurately predict the activities of the excluded retroviral proteases, most notably for the HIV-2 protease with an accuracy of 93% (root mean square error of prediction [RMSEP] = 0.52), and for the HIV-1 protease mutants 86% (RMSEP = 0.65; Figure 2). By contrast, state-of-the-art model building using only the HXB2 HIV-1 protease data failed to give acceptable models (Table 1; see Materials and Methods for further details).

Experimental Validation of CRM

Our approach thus resulted in a statistically well-validated model for the rate of cleavage of peptide substrates by...
retroviral proteases. However, although the model was capable of predicting $k_{cat}/K_m$ values, it could not distinguish cleavable from noncleavable sequences. To allow such predictions, we constructed a cleavability model (CLM) by correlating substrate and protease descriptors and their cross-terms to a vector representing cleavability or noncleavability (Table 2; see Materials and Methods for details). The CLM, which was based on the data for all 61 proteases with all cleavable peptides used for the construction of the CRM as well as an additional large set of noncleavable peptides, almost perfectly classified cleavable and noncleavable substrates (97%) and performed excellently in external predictions of cleavability of new sequences (90.1% accuracy; see Table 2 and Materials and Methods).

Encouraged by these results, we confirmed the predictive power of the models by independent experimental validation. We constructed a virtual peptide library and applied in silico screening to it, first by using the CLM, then followed by the CRM. This process resulted in an unbiased set of 30 novel peptides, selected according to diversity criteria, of which 15 were predicted as cleavable and 15 as noncleavable; the predicted cleavage rates for the cleavable ones ranging over almost three orders of magnitude (see Materials and Methods for details).

The peptides were subsequently synthesized and assayed for their cleavability by the HXB2 HIV-1 protease and three HIV-1 proteases harboring mutations associated with drug resistance. The analysis showed that the CLM could correctly recognize all cleavable substrates as cleavable, and all noncleavable substrates as essentially noncleavable (100% accuracy; Table 3). Moreover, the experimentally determined cleavage rates of the cleavable peptides agreed well with the CRM predicted rates on HXB2 and mutated HIV-1 proteases (68% accuracy; RMSEP = 1.01; Figure 1B). Addition of all experimental data to the CRM further increased CRM

| Parameter | CLM | CRM |
|-----------|-----|-----|
| Number of observations | 2,150 (747 cleavable and 1,403 noncleavable) | 760 |
| Descriptors used | B and C descriptor blocks and $B \times C$, $C \times C$ cross-term descriptor blocks | A, B, and C descriptor blocks; $A \times A$, $A \times B$, $A \times C$, $B \times C$, $C \times C$ cross-term descriptor blocks |
| Model fit ($R^2$) | 0.87 | 0.77 |
| Cross-validation results ($Q^2$) | 0.81 | 0.52 |
| Permutation test results ($R^2$; $Q^2$) | 0.27; −0.16 | 0.31; −0.55 |
| Internal validation | 97% classification accuracy (cutoff, −0.3) | Prediction accuracy 90.1% ± 1.2% |
| External validation | Prediction accuracy 60%–93% for retroviral proteases, RMSEP = 0.52–1.19 $log(k_{cat}/K_m)$ units |

A, B, and C denote descriptor blocks and cross-term descriptor blocks as detailed in the legend to Table 1. For computational details and further explanation of abbreviations, see Materials and Methods.

doi:10.1371/journal.pcbi.0030048.t002

Figure 1. Goodness of Fit and Experimental Validation for the CRM

(A) Observed versus predicted rate of cleavage of 299 substrates by 61 retroviral proteases (in total, 760 protease–substrate combinations) for the CRM, in which all predictions relate to model-building data ($R^2 = 0.77; Q^2 = 0.52$). Each bullet represents naturally occurring and artificially mutated protease forms of HIV-1 (gray), HIV-2 (magenta), AMV (light green), RSV (blue), HTLV-1 (orange), BLV (red), Mo-MuLV (yellow), EIAV (green), and FIV (light blue). (B) A priori prediction of cleavage rates of 15 novel peptides with diverse structures by the CRM (Table 3, numbers 4–18). Shown is the predicted versus experimentally determined cleavage rates by HXB2 (red) and mutant HIV-1 proteases, I84V (blue), L90M (magenta), and I84 L90M (green). The prediction error for the cleavage rates was less than one log($k_{cat}/K_m$) unit for 68% of the protease–substrate pairs; the correlation for the a priori predicted rates versus the experimentally determined rates yielding a correlation coefficient $r = 0.47$ ($p < 0.0001$), as indicated on the panel. The data in (A) is also shown in (B) (gray).

doi:10.1371/journal.pcbi.0030048.g001
Figure 2. External Predictions for Retroviral Proteases by CRMs

Each panel represents the predictions of a model created on the data collected herein (Table S1), but excluded all data for the proteases of one retroviral strain, one at a time, and uses the model to predict the excluded data. Blue bullets correspond to prediction of cleavage activity for new proteases and new substrates (i.e., the cases in which neither the protease nor the peptide was represented in the dataset used in creation of the model). Red bullets correspond to prediction of the cleavage rates for the new proteases only (i.e., the cases in which the peptide, but not the protease, was represented in the dataset used for model creation). Gray dots represent the observed versus computed cleavage rates for each of the respective models (i.e., the data used for model creation).
predictability according to cross-validation; the $Q^2$ increased from 0.52 to 0.54.

Interpretation of the Chemical Effects in Substrates Determining their Cleavage Susceptibility by Retroviral Aspartate Proteases

Analysis of the regression coefficients of the CRM allows analysis of the physicochemical properties of the amino acid residues of both the substrates and the proteases required for catalytic activity. The model verifies the well-known hydrophobic requirement for the $P_3$–$P_3'$ residues of the substrates by the retroviral protease cleavage sites (i.e., the regression coefficients for the $z_1$ terms of the $P_3$–$P_3'$ residues of the substrates are negative, as can be seen from the physicochemical property map derived from the model; Figure 3) [20]. In fact, the map shows that $P_3$ and $P_3'$ can accommodate various amino acids with a preference for hydrophobic residues, which is in perfect agreement with previous findings [13]. Moreover, it is well-known that $\beta$-branched or polar amino acids are not tolerated in the $P_1$ position of retroviral substrates [13]. In the model, this limitation is reflected in that the $z_1$, $z_2$, and $z_4$ terms for the $P_1$ position are most favorable for aromatic amino acids and methionine, while polar or $\beta$-branched residues are disfavored.

Specificity studies for retroviral proteases have found that highly complicated and not easily interpretable interior interactions take place in the substrates [21]. Such interactions become easy to interpret in the model by the cross-terms, however. For example, the large regression coefficients from particular physicochemical properties of the $P_1$–$P_2$, $P_1'$–$P_2$, $P_1$–$P_3$, and $P_1$–$P_4$ residue pairs indicate the presence of interactions for these pairs, while no such interactions take place for the $P_3$–$P_2$ pair according to the model, in accordance with experimental results [22]. Cooperativity between $P_1$–$P_2$, $P_1$–$P_3$, and $P_1$–$P_4$ residue pairs, indicated by the model, has also been shown to be important for specificity features [21]. In earlier specificity studies of retroviral proteases, many series of substrates were used, each of which often differ only by one or two residues, prohibiting a complete analysis of all residues at every subsite [21]. Merging all the available data thus provides a comprehensive picture that reveals important cross-dependencies between several different residue positions in the substrates (i.e., the regression coefficients where particular substrate–substrate cross-terms are significantly large), and demonstrates that a complex interaction network between residues in the substrates is involved in their cleavage (Figure 3).

Interpretation of the Chemical Effects in Aspartate Proteases Involved in their Cleavage of Substrates

In a similar way as above for analysis of substrates, the regression coefficients of the descriptor terms for protease amino acid residues can identify the physicochemical properties of the nonconserved amino acids in the proteases that determine substrate cleavage (e.g., the model reveals that hydrophilic amino acids are preferred at the position corresponding to position 82 in HIV-1 protease to afford a high catalytic activity of the proteases). This is due to the fact that the regression coefficient $z_1$ for position 82 is the largest one and negative. Another example is that hydrophilic, small-size amino acid residues (e.g., Ile or Pro) are preferred at position 81, since both regression coefficients for hydrophobicity ($z_1$) and size ($z_2$) are among the largest and also negative at this position for the model.

To assess a cumulative importance of all physicochemical properties for each protease residue relatively to other residues, we computed and compared the absolute value sums of $z_1$–$z_5$ regression coefficients for each individual position, which is thus a measure representing the overall importance of an amino acid in eliciting chemical effects in the protease when compared to the same measure of other amino acids in the model. Our results from this analysis reveal the most important nonconserved positions involved in catalytic activity of the retroviral proteases (Figure 4A; see Materials and Methods for details). One of the most important amino acid residues shown by the CRM was the threonine of the aspartate proteases' catalytic triad, Asp-Thr(Ser)-Gly (i.e., the T26 residue in HIV-1 protease that is substituted to serine in Rous sarcoma and avian myeloblastosis virus retroviral proteases; Figure S1) [23]. Six further amino acid positions (corresponding to R8, D30, V32, V82, I84, and L90 residues in the HXB2 HIV-1 protease) were also identified as important. These positions agree well with the amino acids known to be associated with high resistance to protease inhibitors [24,25]. The model also identified P81 and N83 amino acid positions, which are known to play a key role in regulation of retroviral protease function [26]. The role of the I64 residue, also indicated by the model, appears to be indirect, as it is located farther way from the substrate cleavage cleft (Figure 4A).

The substrate-protease cross-terms of the CRM were then in a similar fashion used to identify the major cross-dependencies of the protease and substrate amino acids for cleavage activity, which thus reveal the major interaction effects that determine substrate specificity (Figure 4B–4D; see Materials and Methods for details). We then found that $P_3'$ substrate residues form the strongest cross-dependencies with retroviral protease amino acids corresponding to L24, D29, I84, and L97 residues in the HIV-1 protease (Figure 4D). Notwithstanding that D29 directly contributes to the S3' subsite, the effect of residues L24, I84, and L97 distal to the S3' subsite is indirect [13]. Further analysis indicated the importance of direct interactions between the $P_1$ residue and the P81 and V82 protease amino acids (Figure 4C), which form a part of the S1 subsite [13].

The $P_1'$ residue, on the other hand, shows a major indirect interaction with the L90 amino acid (Figure 4C). This is a position for a distantly acting, commonly appearing drug-
| Number | Substrate | Wild-Type Observed Rate of Cleavage, log(k_{cat}/K_m)/log(mM h^{-1}) | Wild-Type Predicted Rate of Cleavage, log(k_{cat}/K_m)/log(mM h^{-1}) | I84V Observed | I84V Predicted | L90M Observed | L90M Predicted | I84V = L90M Observed | I84V = L90M Predicted |
|--------|-----------|---------------------------------------------------------------|---------------------------------------------------------------|---------------|---------------|---------------|---------------|-----------------------|-----------------------|
| 1      | H-R-gle-S-Q-N-Y-P-I-V-Q-lyd-R-OH* | 6.00 ± 0.10 | — | 5.86 ± 0.04 | — | 5.76 ± 0.40 | — | 5.59 ± 0.05 | — |
| 2      | H-R-gle-L-N-F-P-I-S-P-lyd-R-OH | 5.00 ± 0.10 | — | 4.67 ± 0.11 | — | 4.67 ± 0.16 | — | 4.33 ± 0.03 | — |
| 3      | H-R-gle-R-K-I-L-I-F-L-D-G-lyd-R-OH | 5.62 ± 0.19 | — | 5.36 ± 0.05 | — | 5.54 ± 0.07 | — | 5.04 ± 0.07 | — |
| 4      | H-R-gle-Q-G-I-M-L-I-Q-O-lyd-R-OH | 5.96 ± 0.32 | 5.77 | 5.97 ± 0.09 | 5.81 | 4.97 ± 0.14 | 5.89 | 5.84 ± 0.05 | 5.87 |
| 5      | H-R-gle-A-G-I-F | 5.62 ± 0.38 | 5.47 | 5.24 ± 0.02 | 5.54 | 5.48 ± 0.09 | 5.77 | 4.94 ± 0.04 | 5.76 |
| 6      | H-R-gle-R-R-N-F-I-Q-T-lyd-R-OH | 5.59 ± 0.19 | 4.57 | 4.98 ± 0.10 | 4.65 | 5.04 ± 0.06 | 4.90 | 4.47 ± 0.04 | 4.88 |
| 7      | H-R-gle-A-Q-I-F-L-A-V-G-lyd-R-OH | 5.26 ± 0.09 | 5.43 | 4.56 ± 0.09 | 5.45 | 2.99 ± 0.05 | 5.58 | 3.14 ± 0.21 | 5.57 |
| 8      | H-R-gle-P-K-N-Y-F-V-D-lyd-R-OH | 5.10 ± 0.26 | 4.39 | 4.01 ± 0.17 | 4.47 | 4.38 ± 0.34 | 4.66 | 3.83 ± 0.07 | 4.64 |
| 9      | H-R-gle-K-R-A-Y | 5.43 ± 0.36 | 5.72 | 4.52 ± 0.01 | 4.80 | 4.48 ± 0.08 | 4.01 | 4.18 ± 0.15 | 4.00 |
| 10     | H-R-gle-A-E-V-M-L-V-V-S-lyd-R-OH | 4.86 ± 0.24 | 5.67 | 4.44 ± 0.10 | 5.72 | 2.72 ± 0.11 | 5.84 | 4.39 ± 0.25 | 5.83 |
| 11     | H-R-gle-P-R-A-Y | 4.73 ± 0.23 | 5.78 | 3.96 ± 0.39 | 3.86 | 4.20 ± 0.10 | 4.01 | 3.81 ± 0.08 | 3.99 |
| 12     | H-R-gle-A-K-I-L-F | 5.45 ± 0.04 | 5.45 | 4.23 ± 0.09 | 5.51 | 3.59 ± 0.03 | 5.65 | 4.06 ± 0.14 | 5.63 |
| 13     | H-R-gle-S-A-E-Y | 4.46 ± 0.30 | 3.64 | 3.79 ± 0.09 | 3.74 | 3.99 ± 0.03 | 3.91 | 3.26 ± 0.13 | 3.89 |
| 14     | H-R-gle-P-R-G-Y | 3.94 ± 0.04 | 4.01 | 3.55 ± 0.11 | 4.08 | 3.70 ± 0.16 | 4.25 | 3.07 ± 0.04 | 4.24 |
| 15     | H-R-gle-P-K-A-Y | 3.93 ± 0.17 | 3.59 | 3.86 ± 0.05 | 3.66 | 3.96 ± 0.05 | 3.81 | 3.40 ± 0.01 | 3.79 |
| 16     | H-R-gle-P-R-N-Y | 3.92 ± 0.08 | 4.13 | 2.78 ± 0.23 | 4.19 | 3.64 ± 0.08 | 4.39 | 3.17 ± 0.18 | 4.38 |
| 17     | H-R-gle-R-Q-N-F | 3.31 ± 0.39 | 4.37 | 2.96 ± 0.24 | 3.78 | 2.71 ± 0.31 | 3.97 | 2.49 ± 0.10 | 3.95 |
| 18     | H-R-gle-P-K-T-Y | 3.15 ± 0.14 | 3.88 | 2.83 ± 0.12 | 3.95 | 3.04 ± 0.45 | 4.10 | 3.04 ± 0.12 | 4.08 |
| 19     | H-R-gle-R-A-N-W-V-T-R-M-lyd-R-OH | SC<sup>b</sup> | NC | SC | NC | SC | NC | NC | NC |
| 20     | H-R-gle-K-E-N-L-V-T-I-lyd-R-OH | NC | SC | NC | SC | NC | NC | NC | NC |
| 21     | H-R-gle-I-K-A-T-M-A-Q-D-M-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |
| 22     | H-R-gle-R-G-E-L-W-T-M-P-lyd-R-OH | SC | NC | SC | NC | SC | NC | SC | NC |
| 23     | H-R-gle-R-G-R-G-W-A-Q-M-P-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |
| 24     | H-R-gle-K-G-N-L-W-I-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |
| 25     | H-R-gle-R-G-Y-A-T-Q-M-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |
| 26     | H-R-gle-R-G-E-M-V-L-S-M-lyd-R-OH | NC | SC | SC | SC | SC | SC | SC | SC |
| 27     | H-R-gle-R-G-K-Q-N-L-A-A-T-P-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |
| 28     | H-R-gle-R-A-R-M-A-T-M-P-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |
| 29     | H-R-gle-R-G-E-A-V-R-M-lyd-R-OH | SC | NC | SC | NC | SC | NC | SC | NC |
| 30     | H-R-gle-R-E-N-M-L-T-O-lyd-R-OH | SC | NC | SC | NC | SC | NC | SC | NC |
| 31     | H-R-gle-R-K-T-W-A-A-R-P-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |
| 32     | H-R-gle-K-G-E-L-A-T-Q-T-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |
| 33     | H-R-gle-R-G-E-L-P-A-D-V-lyd-R-OH | NC | NC | NC | NC | NC | NC | NC | NC |

Substrate numbers 1–3 represent the natural cleavage sites in the Gag-Pol polyprotein. Substrate numbers 4–33 were chosen for experimental validation of the CLM and CRM.  
<sup>*</sup> denotes a scissile bond.  
<sup>a</sup>lyd indicates L-Lys(DABCYL); see Materials and Methods for details.  
<sup>b</sup>Only a slight cleavage of the substrate was observed, with cleavage rates being far too low to be quantified in terms of k_{cat}/K_m; hence, these substrates were regarded as essentially noncleavable.

doi:10.1371/journal.pcbi.0030048.t003

Retroviral Protease Interaction Maps
Although the P3 amino acid may interact directly with the F53 residue, where mutation to a smaller known at such positions. For example, the Cbz group of the drug-resistant retroviral multi-mutants with the mutations pressure, which compensate a decreased catalytic activity of positions of natural cleavage sites under antiviral drug treatment. Moreover, mutations at the E34 position have been seen in clinical HIV samples after protease inhibitor treatment. Although the P3 amino acid may interact directly with various amino acids in the S1 pocket, our results suggest that the P3 amino acid specificity is determined indirectly by effects arising from the I13 and E34 residues (Figure 4C). This result is in alignment with other reports, where the polymorphic mutation I13V was linked with the mutation of Thr to Ala at the P3 position of the natural cleavage site. Moreover, mutations at the E34 position have been seen in clinical HIV samples after protease inhibitor treatment.

The analysis further demonstrated that the P4 and P4’ residues form a large number of important cross-terms with protease amino acids (Figure 4B–4D). The P4 and P4’ positions can broadly tolerate a variety of amino acids (Figure 3). However, mutations could occur in the P4 and P4’ positions of natural cleavage sites under antiviral drug pressure, which compensate a decreased catalytic activity of drug-resistant retroviral multi-mutants with the mutations depicted in Figure 4B–4D. Indeed, resistance mutations are known at such positions. For example, the Cbz group of the retroviral protease inhibitor TL-3 occupies the S1 subsite and interacts with the F53 residue, where mutation to a smaller resistant mutation, L90M in the HIV-1 protease, which has been observed to increase the cleavage activity of HIV-1 protease for natural substrates mutated in the P4’ position. Although the P3 amino acid may interact directly with various amino acids in the S1 pocket, our results suggest that the P3 amino acid specificity is determined indirectly by effects arising from the I13 and E34 residues (Figure 4C). This result is in alignment with other reports, where the polymorphic mutation I13V was linked with the mutation of Thr to Ala at the P3 position of the natural cleavage site. Moreover, mutations at the E34 position have been seen in clinical HIV samples after protease inhibitor treatment.

The analysis further demonstrated that the P4 and P4’ residues form a large number of important cross-terms with protease amino acids (Figure 4B–4D). The P4 and P4’ positions can broadly tolerate a variety of amino acids (Figure 3). However, mutations could occur in the P4 and P4’ positions of natural cleavage sites under antiviral drug pressure, which compensate a decreased catalytic activity of drug-resistant retroviral multi-mutants with the mutations depicted in Figure 4B–4D. Indeed, resistance mutations are known at such positions. For example, the Cbz group of the retroviral protease inhibitor TL-3 occupies the S1 subsite and interacts with the F53 residue, where mutation to a smaller resistant mutation, L90M in the HIV-1 protease, which has been observed to increase the cleavage activity of HIV-1 protease for natural substrates mutated in the P4’ position. Although the P3 amino acid may interact directly with various amino acids in the S1 pocket, our results suggest that the P3 amino acid specificity is determined indirectly by effects arising from the I13 and E34 residues (Figure 4C). This result is in alignment with other reports, where the polymorphic mutation I13V was linked with the mutation of Thr to Ala at the P3 position of the natural cleavage site. Moreover, mutations at the E34 position have been seen in clinical HIV samples after protease inhibitor treatment.

The analysis further demonstrated that the P4 and P4’ residues form a large number of important cross-terms with protease amino acids (Figure 4B–4D). The P4 and P4’ positions can broadly tolerate a variety of amino acids (Figure 3). However, mutations could occur in the P4 and P4’ positions of natural cleavage sites under antiviral drug pressure, which compensate a decreased catalytic activity of drug-resistant retroviral multi-mutants with the mutations depicted in Figure 4B–4D. Indeed, resistance mutations are known at such positions. For example, the Cbz group of the retroviral protease inhibitor TL-3 occupies the S1 subsite and interacts with the F53 residue, where mutation to a smaller resistant mutation, L90M in the HIV-1 protease, which has been observed to increase the cleavage activity of HIV-1 protease for natural substrates mutated in the P4’ position. Although the P3 amino acid may interact directly with various amino acids in the S1 pocket, our results suggest that the P3 amino acid specificity is determined indirectly by effects arising from the I13 and E34 residues (Figure 4C). This result is in alignment with other reports, where the polymorphic mutation I13V was linked with the mutation of Thr to Ala at the P3 position of the natural cleavage site. Moreover, mutations at the E34 position have been seen in clinical HIV samples after protease inhibitor treatment.

The analysis further demonstrated that the P4 and P4’ residues form a large number of important cross-terms with protease amino acids (Figure 4B–4D). The P4 and P4’ positions can broadly tolerate a variety of amino acids (Figure 3). However, mutations could occur in the P4 and P4’ positions of natural cleavage sites under antiviral drug pressure, which compensate a decreased catalytic activity of drug-resistant retroviral multi-mutants with the mutations depicted in Figure 4B–4D. Indeed, resistance mutations are known at such positions. For example, the Cbz group of the retroviral protease inhibitor TL-3 occupies the S1 subsite and interacts with the F53 residue, where mutation to a smaller
homodimers, the modeling does not allow a distinction between cases

(A) The amino acids are shown in red on the 3-D structure of the HXB2 HIV-1 protease as a template. Because the retroviral proteases are homodimers, the modeling does not allow a distinction between cases extracted from the noncleavable fragments located between the observed cleavage sites by using an eight-residue-long sliding window. Some of the data were generated in-house as described below (see Materials and Methods further below).

B

Description of proteases. The 61 retroviral protease sequences included in the study (Table S2) were aligned using the template shown in Figure S1. A total of 94 amino acids could be fully aligned over all the proteases, but only the positions lacking gaps in all proteases, as well as those being nonconserved, were considered. These then amounted to 85 amino acids, which were described by their five principal physicochemical properties, or “z-scales” [16]. These z-scales roughly represent hydrophobicity ($z_1$), steric properties ($z_2$), polarizability ($z_3$), and polarity and electronic effects of amino acids ($z_4$, $z_5$) (z-scales are the principal components of 26 physicochemical properties of amino acids, which include: molecular weight, van der Waals volume, heat of formation, energy of lowest unoccupied molecular orbital, energy of lowest unoccupied molecular orbital, log P, $\sigma$-polarizability, absolute hydrophobicity, absolute hardness, total molecular surface area, polar molecular surface area, nonpolar molecular surface area, number of hydrogen bond donors, number of hydrogen bond acceptors, indicator of positive charge in the side chain, indicator of negative charge in the side chain, NMR $\delta$-proton shifts at pH 2.7 and 12.5, and seven descriptors representing thin-layer chromatographic mobilities using different stationary and mobile phases [16]).

Thus, every protease was described by $85 \times 5 = 425$ descriptors, which comprised the physicochemical property space information of the series of proteases used herein. It shall be noted that amino acids entirely conserved in a library do not yield any additional information and their importance can therefore not be assessed unless the library is extended by further mutations of these positions.

Description of substrates. We restricted the length of the substrates to octapeptides ($P_1'$-$P_2'$-$P_3'$-$P_4'$-$P_1$-$P_2$-$P_3$-$P_4$), where $P_1$ represents substrate N-terminus amino acid and $P_1'$ represents C-terminus substrate amino acid), since generally only eight amino acid residues are involved in the interaction process with eight subsites ($S_1$-$S_2$-$S_3$-$S_4$-$S_5$-$S_6$-$S_7$-$S_8$) of a retroviral protease, with the cleavage site being between $P_1$ and $P_1'$. Each one of the eight amino acids of the substrates were described by the same five z-scales as above, yielding $8 \times 5 = 40$ total descriptors for each substrate. This comprised the physicochemical space information of the series of substrates used herein.

Description of assay conditions. Descriptors for eight constituents of the experimental assays according to the published data used [32–64] were included in the modeling in order to normalize for the differences in assay conditions. The descriptors used are given in Table S3 and accounted for variations in pH, sodium chloride, 2-mercaptoethanol, EDTA, DMSO, dithiothreitol, nonidet-P40, and glycerol concentrations.

Description of cross-dependencies of proteases, substrates, and assays. The mutual dependencies of protease, substrate, and assay properties were described by cross-terms. These cross-terms were formed by multiplication of any two of the above-described descriptors of proteases, substrates, and assays. To simplify the discussion in the following, the above blocks of descriptors for assays, proteases, and

Figure 4. The Ten Most Important Nonconserved Residues in Retroviral Protease for Substrate Recognition and the Most Important Cross-Dependences of Retroviral Protease and Substrate Amino Acids Identified by Use of the CRM

Materials and Methods

Data and data preprocessing. Data for substrate cleavage by 61 retroviral proteases were collected in an extensive survey that included publicly available data for retroviral proteases during 1990–2005 [32–64]. The survey included proteases from the following viruses: HIV-1, HIV-2, AMV (avian myeloblastosis virus), RSV (Rous sarcoma virus), HTLV-1 (human T cell leukemia virus type 1), BLV (bovine leukemia virus), Mo-MuLV (Moloney murine leukemia virus), EIAV (equine infectious anemia virus), and FIV (feline immunodeficiency virus); its outcome is summarized in Tables S1 and S2. In some cases fully denaturated proteins had been exposed to HIV-1 or HIV-2 proteases [61,62,64]. Noncleavable octapeptides were in these cases extracted from the noncleavable fragments located between the observed cleavage sites using an eight-residue-long sliding window. Some of the data were generated in-house as described below (see Materials and Methods further below).

Description of proteases. The 61 retroviral protease sequences included in the study (Table S2) were aligned using the template shown in Figure S1. A total of 94 amino acids could be fully aligned over all the proteases, but only the positions lacking gaps in all proteases, as well as those being nonconserved, were considered. These then amounted to 85 amino acids, which were described by their five principal physicochemical properties, or “z-scales” [16]. These z-scales roughly represent hydrophobicity ($z_1$), steric properties ($z_2$), polarizability ($z_3$), and polarity and electronic effects of amino acids ($z_4$, $z_5$) (z-scales are the principal components of 26 physicochemical properties of amino acids, which include: molecular weight, van der Waals volume, heat of formation, energy of lowest unoccupied molecular orbital, energy of lowest unoccupied molecular orbital, log P, $\sigma$-polarizability, absolute electro-negativity, absolute hardness, total molecular surface area, polar molecular surface area, nonpolar molecular surface area, number of hydrogen bond donors, number of hydrogen bond acceptors, indicator of positive charge in the side chain, indicator of negative charge in the side chain, NMR $\delta$-proton shifts at pH 2.7 and 12.5, and seven descriptors representing thin-layer chromatographic mobilities using different stationary and mobile phases [16]).

Thus, every protease was described by $85 \times 5 = 425$ descriptors, which comprised the physicochemical property space information of the series of proteases used herein. It shall be noted that amino acids entirely conserved in a library do not yield any additional information and their importance can therefore not be assessed unless the library is extended by further mutations of these positions.

Description of substrates. We restricted the length of the substrates to octapeptides ($P_1'$-$P_2'$-$P_3'$-$P_4'$-$P_1$-$P_2$-$P_3$-$P_4$), where $P_1$ represents substrate N-terminus amino acid and $P_1'$ represents C-terminus substrate amino acid), since generally only eight amino acid residues are involved in the interaction process with eight subsites ($S_1$-$S_2$-$S_3$-$S_4$-$S_5$-$S_6$-$S_7$-$S_8$) of a retroviral protease, with the cleavage site being between $P_1$ and $P_1'$. Each one of the eight amino acids of the substrates were described by the same five z-scales as above, yielding $8 \times 5 = 40$ total descriptors for each substrate. This comprised the physicochemical space information of the series of substrates used herein.

Description of assay conditions. Descriptors for eight constituents of the experimental assays according to the published data used [32–64] were included in the modeling in order to normalize for the differences in assay conditions. The descriptors used are given in Table S3 and accounted for variations in pH, sodium chloride, 2-mercaptoethanol, EDTA, DMSO, dithiothreitol, nonidet-P40, and glycerol concentrations.

Description of cross-dependencies of proteases, substrates, and assays. The mutual dependencies of protease, substrate, and assay properties were described by cross-terms. These cross-terms were formed by multiplication of any two of the above-described descriptors of proteases, substrates, and assays. To simplify the discussion in the following, the above blocks of descriptors for assays, proteases, and

where only one or both of the amino acids of the homoprotein should be assigned as important.

(B) The retroviral protease amino acid residues most important for $P_4$ substrate position (shown in blue).

(C) The retroviral protease amino acid residues most important for the $P_3$ (light blue), $P_2$ (yellow), $P_1$ (red), and $P_1'$ (magenta) substrate positions.

(D) The retroviral protease amino acid residues most important for the $P_4'$ (white) and $P_4''$ (orange) substrate positions (see Materials and Methods for details).

doi:10.1371/journal.pcbi.0030048.g004
substances will be referred to as A, B, and C descriptor blocks, respectively. The cross-terms were then formed by multiplications yielding A × A, A × B, A × C, B × C, and C × C cross-term blocks. Each one of these blocks were in the subsequent modeling used in various combinations, together with the A, B, and C blocks, to demonstrate their respective importance and to find the most suited combination for creation of optimal models. All ordinary protease, substrate, and assay descriptors were mean-centered and scaled to unit variance prior to computation of cross-terms. In addition, we applied block-scaling for each type of descriptors to account for their differences in number and mutual correlation [65]. Block-scaling gives each block a variance square root of N, where N is the number of descriptors in block b. Block-scaling thus gains each variable the variance 1/(N) 1/2. The procedure avoids a situation where large blocks of descriptors mask small ones. (The B × B cross-term block was not formed due to its huge number of descriptors [i.e., 90,100 descriptors]).

Multivariate modeling and data analysis. CRM. All experiments listed in Table S1, where substrate cleavage rates had been determined, were used for the construction of CRM, and comprised 760 observations. Protease, substrate, and assay descriptors, and cross-terms thereof, were used as detailed in Table 2. The preprocessed descriptors (see below) were correlated to measured cleavage rates log(kcat/Km) units by PLS regression modeling using Simca-P+ 10.0 software (Umetrics AB, http://www.umetric.com). In the model building, inclusion of various descriptor blocks were attempted and the data were subjected to PLS regression modeling (see models 1–5 and the single target model [STM] in Table 1 for details) [65]. While models 1–5 utilized all the 760 log(kcat/Km) values obtained from Table S1, the STM comprised only 212 experiments for the HXB2 HIV-1 protease of Table S1. Models were subjected to validation (see below), and model 4 and the CRM were the only ones considered acceptable (R 2 > 0.7 and Q 2 > 0.4) [66]. As CRM also outperformed model 4, it was the one used herein.

For the CRM containing descriptors of substrates, proteases, assays, and their cross-terms as shown in Table 2, the regression equation can be expressed as follows:

\[
y = y + \sum_{a=1}^{A} (\text{coeff}_a \times x_a) + \sum_{b=1}^{B} (\text{coeff}_b \times x_b) + \sum_{c=1}^{C} (\text{coeff}_c \times x_c) + 0.5 \times (C-1) \sum_{c_1=c+1}^{C} (\text{coeff}_c \times x_{c_1})
\]

where A, B, and C represent the number of descriptors in assay, substrate, and protease blocks respectively, a, b, and c correspond to assay, substrate, and protease descriptors respectively, and coeff denotes a coefficient for a corresponding descriptor or a cross-term.

CRM All data listed in Table S1 were considered for the CRM. Assay descriptors were not included. This was because the assay conditions used have only minor effects on substrate cleavability. In some cases, the assay conditions also had not been specified. All in all, the dataset comprised 2,163 peptide–protease combinations. However, if these differed significantly from the others, they were therefore excluded. This resulted in a final dataset with a total of 2,150 observations, which was used for the model creation. Proteases, substrates descriptors, and cross-terms were used for the CRM construction as denoted in Table 2. The descriptors, preprocessed as described below, were correlated to the peptide cleavability (y) by PLS regression modeling using Simca-P+ 10.0 software [65].

For the CRM containing descriptors of substrates, proteases, and their cross-terms as shown in Table 2, the regression equation can be expressed as follows:

\[
y = y + \sum_{a=1}^{A} (\text{coeff}_a \times x_a) + \sum_{b=1}^{B} (\text{coeff}_b \times x_b) + \sum_{c=1}^{C} (\text{coeff}_c \times x_c) + \frac{0.5 x \sum_{i=1}^{n} (\text{coeff}_{i} \times x_i \times x_i)}{N}
\]

where B and C represent the number of descriptors in substrate and protease blocks, respectively, b and c correspond to substrate and protease descriptors, respectively, and coeff denotes a regression coefficient for a corresponding descriptor or a cross-term.

Validation of models. The goodness-of-model fits were quantified by R 2. This unitless fraction indicates the portion of the total variation of the response that is explained by the model and shows how well a model fits the data [65,66]. We also computed the root mean square error of estimation (RMSEE) to determine the internal calculation error within the model:

\[
RMSEE = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \text{y}_{\text{predicted}})^2}{N}}
\]

where y and y predicted denote the observed and calculated rates by the CRM [65]. N denotes the number of calculated observations.

Cross-validation is a method of estimating the accuracy of a regression model. In cross-validation the dataset is divided into several parts (seven were used herein), with each part used to test a model fitted to the remaining parts, resulting in the cross-validated regression coefficient R 2* [67,68], where a higher Q 2* denotes a better modeling ability [66].

In bootstrap validation the dataset is repeatedly and randomly permuted, yielding new dataset samples with replacements from the original dataset [69,70]. New models are then built on permutated data, and R 2, Q 2, and correlation coefficients between original and permuted response values are estimated. Intercept values for R 2* (IR 2) and Q 2 (IQ 2) reflecting R 2 and Q 2 of random response data were computed from repeated random permutations of the data (100 repeats were done herein) [70]. Negative IQ 2 indicates that it is impossible to get predictive models based on random data.

External validation for the CRM was performed by randomly dividing the dataset into two parts (50% and 70%). The smaller part was excluded and predicted based on a model created from the remaining 70% of the data. This procedure was repeated ten times. For each external validation round we calculated the prediction accuracy (i.e., the fraction of correctly classified substrates to cleavable or noncleavable versus all observations included in the test set).

External validation of the CRM was performed by excluding all data for eight retroviral strains one at a time in their entirety, and then predicting the excluded data using models constructed from the remaining data. In the case of HIV-1 proteases, the HXB2 HIV-1 protease and HIV-1 proteases with five artificial stabilizing mutations, QT6K + L33I + L63I + C67A + C95A, were kept in the model, and the external predictions were performed for the remaining 23 drug-resistant HIV-1 mutants. The prediction accuracy for each model was estimated as the fraction of protease–substrate pairs with prediction error < 1.0 log(kcat/Km) to all protease–substrate pairs used for the respective external prediction. This critical threshold was set based on 2-fold RMSEE for the CRM (0.49 log(kcat/Km); Table 1). We also used RMSEP to evaluate model predictive ability for external datasets [65]. RMSEP can be compared with the root mean square error of internal cross-validation (RMSECV), which illustrates the error of predictions within the model [65]. RMSEP was computed as follows:

\[
\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^{N} (y_i - \text{y}_{\text{predicted}})^2}{N}}
\]

where y denotes the observed rate and y predicted the externally predicted rate by the CRM. RMSECV was calculated in an identical fashion, using for y predicted the predicted rates obtained during internal cross-validation of the CRM [65]. N denotes the number of predicted observations.

The correlation coefficient, r, for the experimentally observed versus predicted cleavage rates by the CRM (Figure 1B and Figure 2A) was determined. All modeled, and the statistical significance, p of the correlation was assessed. The p-value obtained is the probability that a correlation this great or greater (in the positive direction only) would be seen if there was no linear relationship between observed
and predicted cleavage rates. An in-house add-in to Excel (Microsoft, http://www.microsoft.com) was used for the test of correlation. All significance tests were one-sided.

Analysis of CRM. All descriptors used for the CRM construction were mean-centered and scaled to unit variance, as described above. This transformation unified the different ranges of descriptor values allowing the creation of the model. The absolute value is of a descriptor’s coefficient, the larger its impact is on the model’s outcome.

To construct the retroviral protease substrate physicochemical fingerprint graphically shown in Figure 3, we analyzed CRM substrate and substrate–substrate cross-term descriptor coefficients. First, we compared the absolute values of the coefficients to find the largest ones. A total of 17 z-scales of the substrates’ amino acid residues were then identified to be highly important and are shown in Figure 3 as a red sphere if its regression coefficient had a positive value, and as a blue sphere if its coefficient was negative. In a separate analysis, we identified important substrate–substrate cross-terms. These are represented in Figure 3 as red lines if the corresponding cross-term coefficient had a positive value, and as blue lines if it was negative.

To determine the most important protease amino acids shown in Figure 4A, we compared the sum of the absolute values of the five z-scale descriptor coefficients for each of the 85 aligned amino acid positions of the proteases. Summation of the coefficients allowed us to simultaneously capture all the physicochemical property effects caused by the amino acids considered. The ten amino acids with the largest sums of their coefficients and consequently the largest contribution on cleavage rate according to the model were the ones depicted in Figure 4A. Figure 4A was produced using the Visual Molecular Dynamics (VMD) program, version 1.8.3 [71].

The model was further analyzed by considering protease–substrate amino acids interactions as described by cross-terms. Every protease–substrate amino acid pair yields 25 cross-terms, as five z-scales of each amino acid multiplied makes 25 cross-terms. To capture the most important protease–substrate interactions, we calculated the sum of the absolute values of 25 cross-term coefficients for each substrate–protease amino acid pair. We then compared all obtained sums and identified the protease–substrate interactions with the largest influence on the model’s outcome. In total, the 20 most important protease–substrate amino acid pairs are presented in Figure 4B–4D. Figure 4B–4D was also produced using VMD [71].

It may be noted that whereas the regression coefficients arising from the substrates only, or the proteases only, relate to the overall activity of all the substrates and all the proteases, respectively, the coefficients of the substrate–protease cross-terms relate to specificity (i.e., the ability of a particular substrate to prefer a particular protease).

In silico substrate screening. The active site of HIV-1 protease accommodates a sequence of eight amino acid residues (P4–P1) of a substrate, and cleaves it between the P1 and P1′ residues. The potential number of substrates consisting of natural amino acids is thereby huge, a large number being feasible to assess. We therefore constructed a smaller library of octapeptide sequences by considering only the natural amino acids that can frequently be found in retroviral protease substrates as follows: for the P1 position, amino acids were R, K, G, or A; for P4: P, Q, A, R, K, G, or E; for P6: N, E, A, G, T, I, L, V, or Y; for P7: F, Y, W, M, or L; for P1′, P, F, A, L, W, or V; for P1′: L, Q, V, A, T, or I; for P3′: D, S, Q, T, M, V, R, or I; and for P7′: T, M, Q, V, P, G, R, or S. This resulted in a virtual library of 6 × 6 × 8 × 5 × 3 × 8 × 8 = 3,317,760 entries.

The library was first screened using the CRM to filter out all noncleavable substrates for the HXB2 HIV-1 protease. We considered a substrate noncleavable if its predicted cleavability parameter was less than −0.3. This resulted in 2,463,379 cleavable sequences (~74% of the initial library). We then used the CRM to predict the actual rate of cleavage for the cleavable octapeptides. From these we chose 15 substrates; seven with a predicted cleavability of more than 4.2 U, and eight with a rate less than 4.2 U. To ensure that peptides were dissimilar, we first randomly selected substrates with predicted log(kcat/Km) > 4.2 U, allowing at most four amino acids to be identical with the corresponding positions of the substrates in the dataset and in previously isolated substrates. For any of the remaining substrates met the requirements, five-amino-acid similarity was allowed. The same procedure was applied for the eight substrates with the predicted log(kcat/Km) < 4.2 U (Table 3, numbers 4–18). Next, we consecutively chose 15 substrates predicted to be noncleavable by HXB2 (Table S3) and predicted cleavable by the CRM, allowing at most four amino acids to be identical at any same positions among all the substrates already selected (including the cleavable substrates already chosen above). If none of the remaining substrates met the require-
numbers 1–35). A typical reaction mixture (total volume 100 μL) contained variable concentrations of peptide substrates in 0.1 M acetic acid and 1.1 M sodium chloride (pH 5.0 was achieved with sodium hydroxide solution) and 35 ng of enzyme. Reaction was conducted at 37 °C for 60 min (cycle time, 60 s, with 5 s shaking after each cycle). Each experiment was repeated at least three times, and the average value was taken as a final result (Table S3). The kinetic data was analyzed by nonlinear fit using the GraFit program and the basic equation for Michaelis–Menten kinetics [73]. The obtained \( k_{\text{cat}}/K_m \) constants were converted into mM·h\(^{-1}\) units for further use.

**Supporting Information**

**Figure S1.** Structural Alignment of Nine Wild-Type Retroviral Proteases

Found at doi:10.1371/journal.pcbi.0030048.s001 (26 KB DOC).

**Table S1.** Summary of the Dataset for Retroviral Proteases Used Herein

References to publications, description of the data considered under the study, and the number of entries each article added to the dataset are shown.

Found at doi:10.1371/journal.pcbi.0030048.s001 (38 KB DOC).

**Table S2.** The 61 Retroviral Proteases from Nine Retroviruses Included in the Study

Found at doi:10.1371/journal.pcbi.0030048.s002 (21 KB DOC).

**References**

1. Rosenberg N, Jolicoeur P (1997) Retroviral pathogenesis. In: Coffin JM, Hughes SH, Varmus H, editors. Retroviruses. Cold Spring Harbor (New York): Cold Spring Harbor Laboratory Press. pp. 475–586.

2. Joint United Nations Programme on HIV/AIDS (2006) 2006 Report on the Global AIDS Epidemic. Executive Summary. Geneva (Switzerland): UNAIDS. 13:

3. Freire E (2002) Designing drugs against heterogeneous targets. Nature Biotechnol 20: 15–16.

4. Lin YC, Beck Z, Lee T, LeVD, Morris GM, et al. (2002) Alteration of substrate and inhibitor specificity of feline immunodeficiency virus protease. J Virol 76: 4710–4720.

5. Beck ZQ, Morris GM, Elder JH (2002) Defining HIV-1 protease substrate selectivity. Curr Drug Targets Infect Disord 2: 37–50.

6. Randolph JT, DeGoey DA (2004) Peptidomimetic inhibitors of HIV protease. Curr Top Med Chem 4: 1079–1095.

7. Clercq ED (2002) Strategies in the design of antiviral drugs. Nature Rev Drug Discov 1: 13–25.

8. Dash C, Kulkarni A, Dunn B, Rao M (2005) Aspartic peptidase inhibitors: Implications in drug development. Curr Rev Biochem Mol Biol 38: 89–119.

9. Ohtaka H, Freire E (2005) Adaptive inhibitors of the HIV-1 protease. Prog Biophys Mol Biol 88: 193–208.

10. Beck ZQ, Lin YC, Elder JH (2001) Molecular basis for the relative substrate specificity of human immunodeficiency virus type 1 and feline immunodeficiency virus proteases. J Virol 75: 9495–9496.

11. Zahuczky G, Boross P, Bagossi P, Emri G, Copeland TD, et al. (2000) Cloning of the bovine leukemia virus protease in Escherichia coli and comparison of its specificity to that of human T-cell leukemia virus protease. Biochim Biophys Acta 1478: 1–8.

12. Lin YC, Beck Z, Morris GM, Olson AJ, Elder JH (2003) Structural basis for distinctions between substrate and inhibitor specificities for feline immunodeficiency virus and human immunodeficiency virus proteases. J Virol 77: 4659–4669.

13. Wlodawer A, Gustchina A (2000) Structural and biochemical studies of retroviral proteases. Biochim Biophys Acta 1477: 16–34.

14. Dunn BM, Goodenow MM, Gustchina A, Wlodawer A (2002) Retroviral proteases. Genome Biol 3: 0061.1–0061.7.

15. Gustchina A, Weber IT (1990) Comparison of inhibitor binding in HIV-1 protease and in non-viral aspartic proteases: The role of the flap. FEBS Lett 269: 269–272.

16. Sandberg M, Eriksson L, Jonsson J, Sjo¨stro¨m M, Wold S (1998) New chemical descriptors relevant for the design of biologically active peptides. J Med Chem 41: 2481–2491.

17. Szeltner Z, Polgar L (1996) Rate-determining steps in HIV-1 protease catalysis. J Biol Chem 271: 32180–32184.

18. Prusis P, Ulhen S, Petrovska R, Lapinsh M, Wikberg JES (2006) Prediction of indirect interactions in proteins. BMC Bioinformatics 7: 167.

19. Wikberg JES, Lapinsh M, Prusis P (2004) Proteochromatics: A tool for modelling the molecular interaction space. In: Kubiniy H, Muller G, editors. Chemogenomics in drug discovery—A medicinal chemistry perspective. Weinheim (Germany): Wiley-VCH. pp. 289–309.

20. Peritt SC, Simsic J, Loeb DD, Everitt L, Hutchinson CA, et al. (1991) Analysis of retroviral protease cleavage sites reveals two types of cleavage sites and the structural requirements of the P1 amino acid. J Virol 65: 14539–14547.

21. Dauber DS, Ziermann R, Parkin N, Maly DJ, Mahrus S, et al. (2002) Altered substrate specificity of drug-resistant human immunodeficiency virus type 1 protease. J Virol 76: 1359–1368.

22. Ridky TW, Cameron CE, Cameron J, Leis J, Copeland T, et al. (1996) Human immunodeficiency virus type 1 protease substrate specificity is limited by interactions between substrate amino acids bound in adjacent enzyme subsites. J Biol Chem 271: 4709–4717.

23. Navia MA, Fitzgerald PM, McKeever BM, Leu CT, Heimbach JC, et al. (1989) Three-dimensional structure of aspartyl protease from human immunodeficiency virus type-1. Nature 337: 615–620.

24. Johnson VA, Brun-Vezinet F, Clotet B, Conway B, Kuritzkes DR, et al. (2005) Update of the drug resistance mutations in HIV-1: 2005. Top HIV Med 13: 51–57.

25. Ho DD, Toyoshima T, Mo H, Kempf DJ, Norbeck D, et al. (1994) Characterization of human immunodeficiency virus type 1 variants with increased resistance to a C2-symmetric protease inhibitor. J Virol 68: 2016–2020.

26. Loeb DD, Swanstrom R, Everitt L, Manchester M, Stamper SE, et al. (1989) Complete mutagenesis of the HIV-1 protease. Nature 340: 397–400.

27. Feher A, Weber IT, Bagossi P, Boross P, Makalingan B, et al. (2002) Effect of sequence polymorphism and drug resistance on two HIV-1 Gag processing sites. Eur J Biochem 269: 4114–4120.

28. Barrie KA, Perez EE, Lamers SL, Farmerie WG, Dunn BM, et al. (1996) Natural variation in HIV-1 protease, Gag p7 and p6, and protease cleavage sites within gag/pol polyproteins: Amino acid substitutions in the absence of protease inhibitors in mothers and children infected by human immunodeficiency virus type 1. Virology 219: 407–416.

29. Wu TD, Schiffer CA, Gonzales MJ, Taylor J, Kantor R, et al. (2003) Mutation pattern and structural correlates in human immunodeficiency virus type 1 protease following different protease inhibitor treatments. J Virol 77: 4856–4847.

30. Buhler B, Lin YC, Morris GS, Olson AJ, Wong CH, et al. (2001) Viral evolution in response to the broad-based retroviral protease inhibitor TL-3. J Virol 75: 9502–9508.

31. Kantor R, Machekano R, Gonzales MJ, Dupnik K, Schapiro JM, et al. (2001) Human immunodeficiency virus reverse transcriptase and protease sequence database: An expanded data model integrating natural language text and sequence analysis programs. Nucleic Acids Res 29: 296–299.

32. Ridky TW, Kikonyogo A, Emri G, Bagossi P, Emri G, Copeland TD, et al. (2000) Drug-resistant HIV-1 proteases identify enzyme residues important for substrate selection and catalytic rate. Biochemistry 37: 13855–13867.

33. Cameron CE, Grinde B, Jacques P, Jenofft J, Leis J, et al. (1995) Comparison of the substrate-binding pockets of the Rous sarcoma virus and human immunodeficiency virus type 1 protease. J Biol Chem 270: 11711–11720.

34. Grinde B, Cameron CE, Leis J, Weber IT, Wlodawer A, et al. (1992) Analysis of substrate interactions of the Rous sarcoma virus wild type and mutant proteases and human immunodeficiency virus type I protease using a set of systematically altered peptide substrates. J Biol Chem 267: 9491–9498.
43. Boross P, Bagossi P, Copeland TD, Oroszlan S, Louis JM, et al. (1999) Effect of substrate specificity on the P2 residue of HIV protease. Biochemistry 38: 123–129.

44. Tozser J, Friedman D, Weber IT, Blaha I, Oroszlan S (1993) Studies on the S4 and S3′ substrates of avian myeloblastosis virus proteinase and lentiviral proteinases. J Biol Chem 268: 7671–7678.

45. Tozser J, Bagossi P, Boross P, Louis JM, Majerova E, et al. (1999) Effect of serine and tyrosine phosphorylation on retroviral protease substrates. Eur J Biochem 265: 423–429.

46. Louis JM, Oroszlan S, Tozser J (1999) Stabilization from autoproteolysis and inhibition of the retroviral protease by a Rous sarcoma virus peptide substance representing the cleavage site between the gag p2 and p10 proteins. J Biol Chem 274: 25735–25741.

47. Tozser J, Bagossi P, Weber IT, Copeland TD, Oroszlan S (1996) Comparative studies on the substrate specificity of avian myeloblastosis virus proteinase and lentiviral proteinases. J Biol Chem 271: 6781–6788.

48. Tozser J, Friedman D, Weber IT, Blaha I, Oroszlan S (1993) Studies on the substrate specificity of the proteinase of equine infectious anemia virus using oligopeptide substrates. Biochemistry 32: 3347–3353.

49. Lin YC, Beck Z, Lee T, Le VD, Morris GM, et al. (2000) Cloning of the bovine leukemia virus proteinase in Escherichia coli and comparison of its specificity to that of human T-cell leukemia virus proteinase. Biochim Biophys Acta 1478: 1–8.

50. Cameron CE, Grinde B, Jentoft J, Leis J, Weber IT, et al. (1992) Kinetic and modeling studies of S3′–S5′ subsites of HIV-1 protease. In: Biochemistry of HIV-1 protease. Adv Exp Med Biol 279: 179–183.

51. Mahalingam B, Louis JM, Reed CC, Adomat JM, Krouse J, et al. (1999) Atomic detail modeling studies of drug resistant mutants of HIV-1 protease. Eur J Biochem 265: 423–429.

52. Tozser J, Friedman D, Weber IT, Blaha I, Oroszlan S (1993) Studies on the substrate specificity of the protease of equine infectious anemia virus using oligopeptide substrates. Biochemistry 32: 3347–3353.

53. Lin YC, Beck Z, Lee T, Le VD, Morris GM, et al. (2000) Cloning of the bovine leukemia virus proteinase in Escherichia coli and comparison of its specificity to that of human T-cell leukemia virus proteinase. Biochim Biophys Acta 1478: 1–8.

54. Cameron CE, Jentoft J, Leis J, Weber IT, Wlodawer A, et al. (1992) Mechanism of inhibition of the retroviral protease by a Rous sarcoma virus peptide substrate representing the cleavage site between the gag p2 and p10 proteins. J Biol Chem 267: 25735–25741.

55. Tozser J, Bagossi P, Weber IT, Copeland TD, Oroszlan S (1996) Comparative studies on the substrate specificity of avian myeloblastosis virus proteinase and lentiviral proteinases. J Biol Chem 271: 6781–6788.

56. Tozser J, Friedman D, Weber IT, Blaha I, Oroszlan S (1993) Studies on the substrate specificity of the proteinase of equine infectious anemia virus using oligopeptide substrates. Biochemistry 32: 3347–3353.

57. Lin YC, Beck Z, Lee T, Le VD, Morris GM, et al. (2000) Cloning of the bovine leukemia virus proteinase in Escherichia coli and comparison of its specificity to that of human T-cell leukemia virus proteinase. Biochim Biophys Acta 1478: 1–8.

58. Cameron CE, Jentoft J, Leis J, Weber IT, Wlodawer A, et al. (1992) Mechanism of inhibition of the retroviral protease by a Rous sarcoma virus peptide substrate representing the cleavage site between the gag p2 and p10 proteins. J Biol Chem 267: 25735–25741.

59. Tozser J, Bagossi P, Weber IT, Copeland TD, Oroszlan S (1996) Comparative studies on the substrate specificity of avian myeloblastosis virus proteinase and lentiviral proteinases. J Biol Chem 271: 6781–6788.

60. Tozser J, Friedman D, Weber IT, Blaha I, Oroszlan S (1993) Studies on the substrate specificity of the protease of equine infectious anemia virus using oligopeptide substrates. Biochemistry 32: 3347–3353.

61. Lin YC, Beck Z, Lee T, Le VD, Morris GM, et al. (2000) Cloning of the bovine leukemia virus proteinase in Escherichia coli and comparison of its specificity to that of human T-cell leukemia virus proteinase. Biochim Biophys Acta 1478: 1–8.

62. Tozser J, Friedman D, Weber IT, Blaha I, Oroszlan S (1993) Studies on the substrate specificity of the protease of equine infectious anemia virus using oligopeptide substrates. Biochemistry 32: 3347–3353.

63. Lin YC, Beck Z, Lee T, Le VD, Morris GM, et al. (2000) Cloning of the bovine leukemia virus proteinase in Escherichia coli and comparison of its specificity to that of human T-cell leukemia virus proteinase. Biochim Biophys Acta 1478: 1–8.

64. Chou KC, Zhang CT (1993) Studies on the specificity of HIV protease: An artificial neural network model for predicting HIV-1 protease cleavage sites in protein. Adv Exp Med Biol 279: 27148–27152.

65. Umetrics AB (2001) Simca-P 9 User Guide and Tutorial. Umeå (Sweden).

66. Lundstedt T, Seiffert E, Abramo L, Thelin B, Nystrom A, et al. (1998) Experimental design and optimization. Chemometr Intell Lab Syst 42: 3–40.

67. Wold S (1978) Cross-validatory estimation of the number of components in factor and principal components models. Technometrics 20: 397–405.

68. Wakeling IN, Morris JJ (1995) A test of significance for partial least squares regression. J Chemometr 7: 291–304.

69. Efron B (1987) Better bootstrap confidence intervals and approximations. J Am Stat Assoc 82: 171–183.

70. Prusis P, Lundstedt T, Wikberg JES (2002) Proteo-chemometrics analysis of MSH peptide binding to melanocortin receptors. Protein Eng, 15: 305–311.

71. Humphrey W, Dalke A, Schulten K (1996) VMD: Visual molecular dynamics. J Mol Graphics 14: 35–61.

72. Danielson H, Lindgren MT, Markgren PO, Nillroth U (1998) Investigation of substrate and inhibitor specificity of feline immunodeficiency virus protease using oligopeptide substrates. Biochemistry 32: 3347–3353.

73. Tozser J, Friedman D, Weber IT, Blaha I, Oroszlan S (1993) Studies on the substrate specificity of the protease of equine infectious anemia virus using oligopeptide substrates. Biochemistry 32: 3347–3353.

74. Cameron CE, Jentoft J, Leis J, Weber IT, Wlodawer A, et al. (1992) Mechanism of inhibition of the retroviral protease by a Rous sarcoma virus peptide substrate representing the cleavage site between the gag p2 and p10 proteins. J Biol Chem 267: 25735–25741.