Bioorganic & Medicinal Chemistry 17 (2009) 5229–5237

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

journal homepage: www.elsevier.com/locate/bmc

Proteochemometrics mapping of the interaction space for retroviral proteases and their substrates

Aleksejs Kontijevskis [a,b], Ramona Petrovska [a], Sviatlana Yahorava [a], Jan Komorowski [b], Jarl E. S. Wikberg [a,*]

[a] Department of Pharmaceutical Biosciences, Uppsala University, Husargatan 3, SE-75124, Uppsala, Sweden
[b] Linnaeus Centre for Bioinformatics, Uppsala University, Husargatan 3, SE-75124, Uppsala, Sweden

ARTICLE INFO

Article history:
Received 23 November 2008
Revised 1 April 2009
Accepted 17 May 2009
Available online 23 May 2009

Keywords:
Retroviral proteases
Proteochemometrics
Molecular recognition
Resistance
HIV-1 protease inhibitors

ABSTRACT

Understanding the complex interactions of retroviral proteases with their ligands is an important scientific challenge in efforts to achieve control of retroviral infections. Development of drug resistance because of high mutation rates and extensive polymorphisms causes major problems in treating the deadly diseases these viruses cause, and prompts efforts to identify new strategies. Here we report a comprehensive analysis of the interaction of 63 retroviral proteases from nine different viral species with their substrates and inhibitors based on publicly available data from the past 17 years of retroviral research. By correlating physico-chemical descriptions of retroviral proteases and substrates to their biological activities we constructed a highly statistically valid 'proteochemometric' model for the interactome of retroviral proteases. Analysis of the model indicated amino acid positions in retroviral proteases with the highest influence on ligand activity and revealed general physicochemical properties essential for tight binding of substrates across multiple retroviral proteases. Hexapeptide inhibitors developed based on the discovered general properties effectively inhibited HIV-1 proteases in vitro, and some exhibited uniformly high inhibitory activity against all HIV-1 proteases mutants evaluated. A generalized proteochemometric model for retroviral proteases interactome has been created and analysed in this study. Our results demonstrate the feasibility of using the developed general strategy in the design of inhibitory peptides that can potentially serve as templates for drug resistance-improved HIV retardants.

1. Introduction

The ability of HIV to evade antiviral drugs remains a major problem in its control. Proteases play key roles in the life cycles of retroviruses by processing the viral polyproteins into the structural and functional elements that assemble into the infectious virions.[1] Inhibition of a retroviral protease arrests production of mature infectious viral particles and prevents infection of new host cells.[1] Retroviral proteases have therefore become important targets for drugs aimed at treating diseases caused by retroviruses, such as AIDS, leukemia, and myelopathy.[2,3] In particular, many anti-HIV drugs have been developed, but the fast emergence of drug-resistant variants heavily compromises their effectiveness.[1] Because of the extremely high mutation rate in HIV-1, its protease exhibits amino acid polymorphisms even in the absence of the selective pressure of antiretroviral therapy.[1,4] These polymorphisms are found in at least 49 positions of the 99 residues in the HIV-1 protease sequence,[1,4] and even mutations associated with drug resistance are prevalent in drug-naïve strains.[5–7] The sequence differences for subtypes of HIV-1 proteases can be as high as 30%[5] and range from 10% to 70% among the proteases from the entire retroviral class.[8] All of these genetic polymorphisms and the diversity of the retroviral proteases affect their properties and their ability to bind inhibitors.[8]

Most of the currently used HIV-1 protease inhibitors were developed based on the lock-and-key paradigm to fit tightly into the binding cavity of a 'wild-type' (such as the HXB2 strain) HIV-1 protease.[4] The majority of these inhibitors can be generically classified as peptidomimetics in structure, sharing a common hydroxyethylene or hydroxyethylamine core element in place of a normal scissile peptide bond.[9] Because such inhibitors are quite constrained, they lack adaptability to target variations; therefore, it is not surprising that they do not perform well against many of the variants of HIV proteases.[4,10] Substrates for the retroviral proteases, on the other hand, are flexible molecules that can adapt and bind to many protease mutants. Protease substrates with low Michaelis constants, K_m, over multiple retroviral proteases might serve as templates for the design of adaptive peptidomimetic inhibitors that show inhibitory activity across multiple resistance mutations.[11] However, understanding the interactions of substrates with HIV proteases is a complex issue, and it is difficult

* Corresponding author. Tel.: +46 184714238; fax: +46 18559718.
E-mail address: Jarl.Wikberg@farmbio.uu.se (J.E.S. Wikberg).

0968-0896/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2009.05.045

indeed to predict which sequence of amino acids would produce a substrate with low K_m values over multiple protease mutants.

In the present study we used a powerful chemo-bioinformatics approach for protein–ligand interaction space analysis, proteochemometrics,[12–19] to model substrate affinity across multiple retroviral proteases using substrate K_m values. We subsequently used the developed model in the design of inhibitory peptides with affinity for both drug-sensitive and drug-resistant HIV-1 protease variants, and we can now report that the approach successfully produced inhibitory peptides with simultaneous, uniformly high inhibitory activity against drug-sensitive and drug-resistant HIV-1 proteases.

2. Results

2.1. Generalized K_m model for retroviral proteases

To construct a generalized K_m model (GKM) for retroviral proteases, we combined the publicly available assay data for multiple proteases from the last 17 years of retrovirus research into a single data set (Supplementary Table 1). Because retroviral proteases are intrinsically dynamic proteins that change their structures with binding of the substrates, we described each structurally aligned amino acid of the 63 retroviral proteases by their principal physicochemical properties using so called z-scales (termed here 'ordinary protease descriptors'), rather than using fixed protein 3D structures for description.[20,21] In the same way, we described the substrates by considering the principal physicochemical properties of every amino acid of the octapeptide sequence spanning the P_4 to P'_4 position (termed here 'ordinary peptide descriptors'). The ordinary descriptors of the retroviral protease and substrates were then correlated to the experimentally determined K_m values using partial least squares (PLS) regression modeling (model M1 in Table 1). The kinetics of retroviral proteases is dependent on the experimental assay constituents, such as pH and salt concentration.[22] To account for differences in assay conditions, we introduced assay descriptors into the modeling, which significantly improved the model's predictability according to cross-validation (Q^2_{mean} increased by 0.03 in model M2, Table 1; Wilcoxon signed-rank test for model M1 Q^2_{mean} versus model M2 Q^2_{mean} yielded $p < 0.0001$).

To capture the complex processes of substrate recognition involving specific covalent and non-covalent interactions, we introduced 'interaction' terms into the multivariate modeling. Interaction terms are produced by multiplication of any two descriptors and represent approximations of the non-linear parts of molecular interaction effects. The model M2 was then improved by adding different interaction term descriptor blocks, one at a time (models M3–M8, Table 1). Addition of the substrate × substrate, substrate × assay, or substrate × protease descriptor blocks significantly increased the model predictability (Q^2_{mean} difference is statistically significant for models M3, M4, and M5 versus M2 at $p < 0.0001$, according to Wilcoxon signed-rank test). Therefore these descriptor blocks were finally merged with the descriptors used in the M2 model, which resulted in a statistically valid generalized K_m model, GKM, for retroviral proteases (Table 1; see also Eq. 1 in Methods).

The capacity of GKM to predict K_m values for different retroviral proteases and their mutants was then assessed by 'leave-one-protease-out' validation. This evaluation was done by excluding data for one type of virus at a time, creating new models based on the remaining data, and then predicting the excluded data from the new models created (Fig. 1; see Section 4 for details). According to the validation, the log (K_m) values for the excluded proteases and mutants were predicted with reasonable accuracy; the root mean square error of prediction (RMSEP) ranged from 0.41 to 0.62 log (K_m) units, which is similar to the root mean square error of estimation (RMSEE) of GKM itself (Fig. 1; Table 1). In particular, GKM predicted accurately the activities for HIV-1 protease mutants associated with drug-resistance, as well as activities for the HIV-2 protease (Fig. 1A and B). Thus, the leave-one-protease-out validation demonstrated the capacity of the GKM to perform activity predictions for interaction of new proteases with substrates.

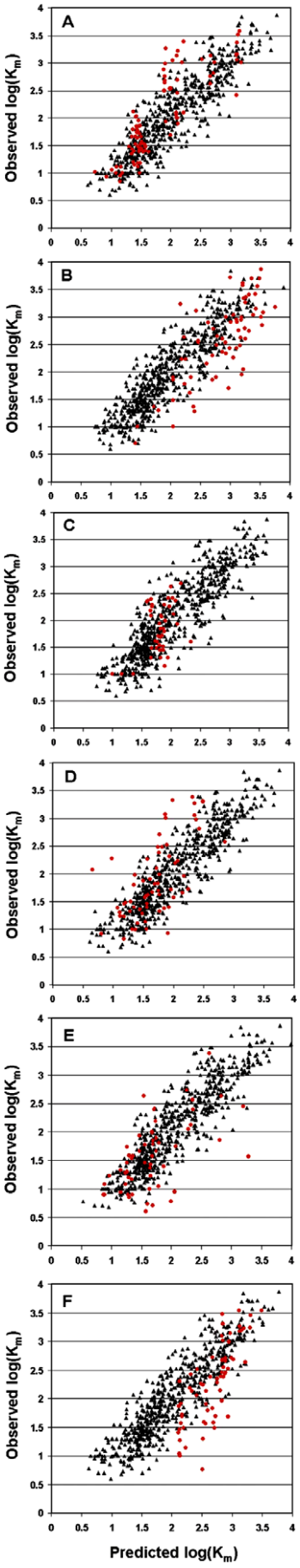
The validity of GKM was further probed by independent external validation. The external dataset comprised experimentally determined K_m values for 15 substrates with diverse amino acid sequences tested on four HIV-1 proteases (HXB2 strain protease and three drug-resistant mutants, I84V, L90M, and I84V+L90M) (Supplementary Table 2, numbers 4–18). The results from the external validation clearly demonstrated that the GKM accurately predicted log (K_m) values for the new substrates and proteases (Fig. 2; Supplementary Table 2).

Quantitative structure–activity relationships (QSAR) modeling is a commonly used modeling approach that aims to correlate descriptors of molecules to their biological activities, and which can consider activities at only one target at a time.[23,24] We here compared the generalized modeling approach, proteochemometrics,

Table 1
Validation of generalized K_m and HIV-1 single-target protease models

Models	R^2	Q^2_{mean} ± SE	RMSEE, log (K_m)	iR^2	iQ^2	NC	Descriptor blocks
M1	0.59	0.47 ± 0.01	0.49	0.05	−0.23	6	Sub, PR
M2	0.60	0.50 ± 0.01	0.47	0.06	−0.25	7	Sub, PR, A
M3	0.64	0.52 ± 0.01	0.44	0.16	−0.21	7	Sub, PR, A, Sub × PR
M4	0.63	0.52 ± 0.01	0.44	0.14	−0.21	6	Sub, PR, A, Sub × Sub
M5	0.69	0.55 ± 0.01	0.41	0.15	−0.36	8	Sub, PR, A, Sub × A
M6	0.58	0.48 ± 0.01	0.47	0.06	−0.21	6	Sub, PR, A, PR × PR
M7	0.59	0.47 ± 0.01	0.47	0.06	−0.24	6	Sub, PR, A, PR × A
M8	0.61	0.49 ± 0.01	0.46	0.07	−0.34	8	Sub, PR, A, A × A
GKM	0.85	0.62 ± 0.02	0.29	0.37	−0.77	12	Sub, PR, A, Sub PR, Sub A, Sub × Sub
STM1	0.44	0.36 ± 0.01	0.53	0.08	−0.13	2	Sub
STM2	0.49	0.38 ± 0.01	0.50	0.11	−0.16	3	Sub, A
STM3	0.59	0.44 ± 0.01	0.45	0.24	−0.09	3	Sub, A, Sub × Sub
STM4	0.64	0.43 ± 0.02	0.42	0.25	−0.25	4	Sub, A, Sub × A
STM5	0.46	0.35 ± 0.01	0.52	0.09	−0.20	3	Sub, A, A × A
STM6	0.82	0.51 ± 0.02	0.30	0.50	−0.42	6	Sub, A, Sub × Sub, Sub × A

Generalized K_m (M1–M8 and GKM) and single-target HIV-1 protease (STM1–STM6) models were developed by inclusion of different descriptor blocks in various combinations. The descriptor blocks were as follows: A, assay constituents descriptor block; PR, protease descriptor block; Sub, substrate descriptor block; and A × A, PR × A, Sub × A, Sub × Sub, Sub × PR, and PR × PR represent the interaction term blocks formed from respective ordinary descriptor blocks. Models M1–M8 and GKM were constructed using all data for retroviral proteases, and single-target models (STM1–STM6) used only data for a 'wild-type' HXB2 strain HIV-1 protease. SE, standard error of the Q^2_{mean}. NC indicates the number of significant PLS components used in the model construction.



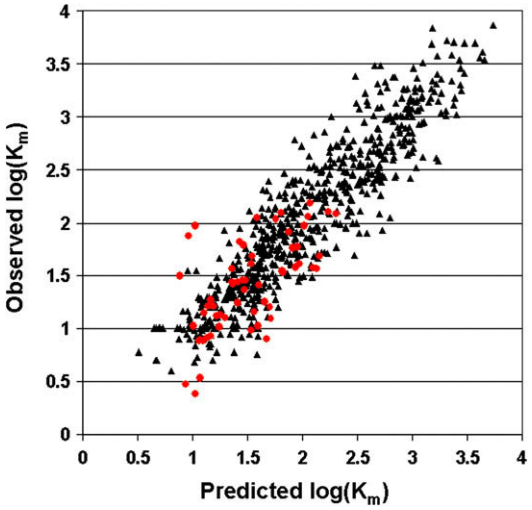


Figure 2. Experimental validation of the generalized K_m model. Red bullets represent *a priori* predictions of K_m constants for 15 peptides with diverse structures for HXB2 HIV-1 protease and three mutant HIV-1 proteases (I84V, L90M, and I84V + L90M) by GKM, ($r = 0.63$, $p < 0.0001$, RMSEP = 0.36) (*Supplementary* Table 2, Nr. 4–18). Black triangles show the fit of GKM for all model-building data.

to a single-target quantitative structure–activity relationship modeling approach using the data for the HXB2 HIV-1 protease among the collected data only. However, the QSAR models for HXB2 HIV-1 protease were statistically invalid or overfitted (i.e., $R^2 < 0.7$ or $Q^2 < 0.4$ for models STM1–STM5 and $iR^2 > 0.4$ for the model STM6; Table 1). Thus, the inclusion of many proteases in one uniform proteochemometric model markedly improves GKM performance and provides essential information about the interaction space of retroviral proteases, that is, the information that the single target modeling approach is entirely lacking.

2.2. Analysis and interpretation of the generalized K_m model

We analyzed the model's regression coefficients to assess the relative importance of descriptors and cross-terms in GKM explaining the Michaelis constant K_m for any protease-substrate-assay combination (see Eq. 1 in Methods). We first determined the non-conserved protease positions having the largest influence on K_m. This determination was made by comparing sums of absolute values of z_1–z_5 regression coefficients for ordinary protease amino acid descriptors of each aligned amino acid position. The positions corresponding to amino acids R8, V32, L33, I64, P81, V82, N83, and I84 in the HXB2 HIV-1 protease showed the largest sums, suggesting that these are the positions playing major roles in protease influence on the K_m values of the native substrates (Table 2). These amino acids overlap between the positions we found

Figure 1. External validation of the generalized K_m model (GKM). Each panel shows the predictions of a model created based on the data collected in the current work, but excluding all data for proteases of one retroviral strain and using the model created on the remaining data to predict the excluded data. Red bullets represent observed versus predicted K_m for the excluded proteases. Black triangles represent observed versus computed K_m for the respective GKM model-building data. Panels A–F represent 'leave-one-protease-out' predictions for wild-type, naturally occurring, and artificially mutated proteases as follows: (A) HIV-1 ($r = 0.85$, $p < 0.0001$, RMSEP = 0.45); (B) HIV-2 ($r = 0.77$, $p < 0.0001$, RMSEP = 0.55); (C) HTLV-1 ($r = 0.44$, $p = 0.0017$, RMSEP = 0.41); (D) AMV ($r = 0.65$, $p < 0.0001$, RMSEP = 0.56); (E) RSV ($r = 0.57$, $p < 0.0001$, RMSEP = 0.54), and (F) Mo-MuLV ($r = 0.77$, $p < 0.0001$, RMSEP = 0.62) proteases. For panel A the data excluded were 23 HIV-1 proteases holding mutations associated with drug resistance.

Table 2
Descriptors of highest importance in the GKM model

Descriptor block	Important descriptors and positions
Substrate	P_3z_1, P'_2z_1, P'_3z_1, P_2z_1, P'_1z_5, P_2z_3, P'_1z_3
Protease	R8, I84, V32, L33, P81, V82, I64, N83
Assay	Dimethyl sulfoxide, pH, sodium chloride
Substrate × substrate	$P_3z_2 \times P'_1z_5$, $P_3z_5 \times P_1z_1$, $P_3z_2 \times P_1z_1$, $P_4z_1 \times P'_1z_5$, $P_3z_5 \times P_1z_4$
Substrate × protease	$P_3 \times$ T31, $P_2 \times$ L90, $P_3 \times$ I13, $P_3 \times$ I85, $P_3 \times$ K20, $P_3 \times$ 72, $P_3 \times$ V32, $P_3 \times$ L33, P'_3 V32, $P_2 \times$ P81

Shown are the descriptors with the largest influence on K_m within each descriptor block, in decreasing order of importance. Amino acid positions for proteases correspond to amino acid positions in the HXB2 HIV-1 protease.

earlier to be associated with the maintenance of high cleavage rates (k_{cat}/K_m) of substrates among the aspartyl proteases (that is, R8, V32, V82, I64, P81, N83, and I84).[19] Mutations in these positions are also known to be associated with resistance to HIV-1 protease inhibitors.[25,26] Although some other conserved amino acids in the active sites of the retroviral proteases may have utmost importance for substrate cleavage and binding, the lack of variations in these positions in the present data-set prohibited us to evaluate their importance relatively to other non-conserved amino acid positions.

Substrate–protease interaction terms were then analyzed in a similar way to uncover the substrate–protease inter-dependencies having the highest impact on K_m (see Section 4 for details). The results suggested that substrate P_3 residues establish important direct or indirect interactions with many protease amino acids (Table 2). This finding is in alignment with earlier reports characterizing the S_3 and S'_3 sub-sites in the retroviral proteases as being large and able to tolerate amino acids of different types and sizes.[6,27] Our result suggests that mutations or polymorphic changes directed to these protease positions (Table 2) can be associated with a compensatory change in substrate P_3 position to restore a reduced fitness to the enzyme. Additionally, we found that the substrate P_2 position shows important cross-dependencies with L90 and P81 protease positions, while the P'_3 position shows an important cross-dependency with the V32 position (Table 2).

The GKM model achieves further high impact from the assay descriptors representing pH, salt, and dimethyl sulfoxide concentrations (Table 2), results that also are in agreement with earlier findings.[28–32]

The coefficients for ordinary substrate descriptors were finally analyzed to reveal the physicochemical properties of the substrates

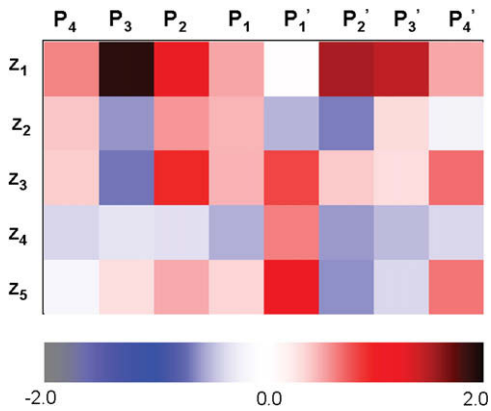


Figure 3. PLS coefficients of the substrate principal physicochemical properties determined by the GKM in color coding. The intensity of the color indicates the value and the direction of the coefficient according to the spectrum. The figure demonstrates the physicochemical requirements that a substrate should possess to afford low K_m over retroviral proteases in general.

of general importance for maintaining K_m among the retroviral protease class (see Fig. 3 and Eq. 1 and associated text in Section 4). As seen from Figure 3, the coefficients for the z_1-scale descriptors were the largest and most positive for amino acids at substrate positions P_3, P_2, P'_2, and P'_3, suggesting that hydrophobicity is a critical feature at these positions for affording tight binding of peptides with many retroviral proteases; that is, hydrophobic amino acids have large negative values for their z_1-scales and the product coeff × z_1 will then be large and negative, reducing the overall K_m value for a peptide according to the model's prediction (Eq. 1 in Section 4).

Moreover, the model predicts that amino acids with large negative z_3-scales (reflecting polarizability) are also favorable at positions P_2 and P'_1. Additionally, electronic effects by amino acids at the P'_1 position significantly influence K_m because of the large coefficient for the P'_1z_5 descriptor (Fig. 3). The results further suggest that hydrophobicity (z_1-scale) at the P_1 and P'_1 positions has only a minor impact on K_m, whereas it substantially influences cleavage rate k_{cat}/K_m of the substrates.[11,19] The result is particularly interesting because flexible neutral or hydrophilic amino acids could be placed at the peptide P_1 and P'_1 positions, and the peptides could possibly maintain their ability to bind efficiently to multiple proteases while at the same time themselves becoming uncleavable; thus, they would be multiple protease selective inhibitors (see Eq. 2 under Section 4). Further analysis indicates that amino acids in the P_4 and P'_4 positions only slightly affect substrate K_m; the PLS coefficients of the GKM at these positions were small. This finding suggests that the length of the active octapeptides might be reduced to six amino acids without significantly affecting peptide binding.

2.3. Use of GKM for design of inhibitory peptides

The above results provide general insights into the physicochemical determinants affecting the binding of substrates to retroviral proteases. These determinants are of direct use for selecting coding and non-coding amino acids for each substrate position to result in a peptide that would likely bind tightly to multiple proteases. Thus, for example, the amino acids Lys, Glu, or Thr have negative values for their z_3 and z_5 scales and would thus be favorable selections for the P'_1 position (see a full table of z-scales for coding and non-coding amino acids in Ref. [20]). The P_2 position should favor amino acids such as Ile, Val, Met, or cyclohexylalanine because their z_1 and z_3-scales are negative (Fig. 3).[20] We elected to construct a virtual hexapeptide library based on sets of amino acids for each position, those sets that the GKM indicated should be favorable as a low-K_m peptide for retroviral proteases (see Section 4). To ascertain that chosen peptides were uncleavable, the library was first virtually screened on the cleavability model developed for the retroviral substrates in Ref. [19]. The peptides were then virtually screened on GKM to find peptides with a predicted $K_m < 3$ μM. Of the total 4154 sequences these virtual screenings identified, we selected 10 (Table 3, numbers 1–10) according to set diversity criteria (see Section 4), synthesized them, and tested their inhibitory activity on the HXB2 HIV-1 protease and three drug-resistant HIV-1 protease mutants (I84V, L90M, and I84V+L90M) which were available in our lab. We also evaluated 10 other hexapeptides (Table 3, numbers 11–20) available from earlier, unrelated studies (unpublished data). Our results demonstrated that all 10 of the GKM-predicted peptides with K_m values in the low micromolar range also inhibited the HXB2 HIV-1 protease in the low micromolar K_i range. Moreover, the majority of these hexapeptides also inhibited the mutant proteases with K_i values in the low micromolar range (Table 3). On the other hand, the 10 hexapeptides with GKM-predicted K_m values in the high micromolar range showed no inhibition activity against HIV-1 proteases (Table 3).

Table 3
Experimentally determined K_i values for HIV-1 protease inhibitors

No.	Hexapeptide sequences	Observed K_i, μM ± SD				GKM predicted K_m (μM)			
		Wild-type	L90M	I84V	I84V + L90M	Wild-type	L90M	I84V	I84V + L90M
1	Nal-Cha-Met-Glu-Phg-Cha	5.5 ± 1.4	2.4 ± 0.5	3.9 ± 0.4	2.7 ± 0.4	0.64	0.58	0.86	0.77
2	Nal-Val-Met-Har-Tyr-Nva	5.4 ± 0.6	6.9 ± 2.0	19.9 ± 6.9	7.5 ± 0.3	1.87	1.57	2.33	1.97
3	Hph-Nva-Dap-Har-Nal-Nva	5.2 ± 2.0	23.6 ± 8.8	13.3 ± 5.0	15.7 ± 6.0	0.40	0.35	0.54	0.47
4	Nal-Cha-Dap-Glu-Nal-Nva	12.0 ± 2.5	9.7 ± 0.4	20.7 ± 0.8	20.8 ± 4.1	0.23	0.21	0.32	0.29
5	Cph-Cha-Dap-Glu-Bph-Leu	5.5 ± 1.4	15.7 ± 5.9	37.7 ± 2.0	34.6 ± 4.2	0.66	0.59	0.89	0.79
6	Nal-Cha-Lys-Aph-Btr-Nle	24.1 ± 8.9	9.2 ± 2.9	31.6 ± 4.6	56.0 ± 11.5	0.90	0.73	1.15	0.94
7	Nal-Nle-Dab-Glu-Hph-Nva	22.0 ± 4.1	16.7 ± 6.5	NI	NI	0.80	0.72	1.06	0.96
8	Cph-Cha-Dab-Glu-Phe-Cha	3.8 ± 0.1	31.1 ± 1.9	NI	NI	1.68	1.52	2.24	2.03
9	Hph-Cha-Dap-Har-Phe-Leu	4.6 ± 1.1	26.5 ± 2.7	NI	NI	0.80	0.70	1.06	0.93
10	Hph-Cha-Dab-Arg-Hph-Cha	2.0 ± 0.5	47.4 ± 0.4	NI	NI	0.59	0.51	0.79	0.69
11	Thr-Ala-Ala-Gly-Arg-Thr	NI	NI	NI	NI	124	108	158	138
12	Tyr-Ala-Thr-Pro-Gly-Thr	NI	NI	NI	NI	70	60	93	80
13	Ser-Val-Arg-Cys-Ser-Trp	NI	NI	NI	NI	143	129	191	172
14	Gly-Thr-Ala-Tyr-Ser-Cys	NI	NI	NI	NI	676	608	930	838
15	Asp-Gly-Gly-Ala-Leu-Ser	NI	NI	NI	NI	57	53	75	69
16	Pro-Tyr-Ala-Gly-Ala-Gln	NI	NI	NI	NI	67	66	87	86
17	Val-Arg-Ser-Ala-His-Ile	NI	NI	NI	NI	63	64	89	91
18	Thr-Asn-Thr-Thr-Ala-Asp	NI	NI	NI	NI	108	95	141	124
19	Ala-His-Tyr-Ala-Thr-His	NI	NI	NI	NI	72	79	94	102
20	Ser-Pro-Ala-Thr-Glu-Ala	NI	NI	NI	NI	81	76	106	100

Wild-type denotes HXB2 HIV-1 protease; I84V, L90M, and I84V + L90M are the corresponding mutant HXB2 forms. Under 'Observed' are the experimentally determined K_i values of hexapeptides towards these proteases. Under 'GKM predicted' are the corresponding K_m values predicted by GKM. Peptides 1–10 were unacetylated in their N-terminus while peptides 11–20 were acetylated. All peptides had an amide group in their peptide C-terminus. SD, standard deviation. NI, no inhibition observed at 100 μM of the hexapeptide. (For abbreviations of artificial amino acids see **Section** 4).

Furthermore, our results clearly indicated a strong, statistically significant correlation between GKM-predicted log (K_m) and the observed log (K_i) for all 20 hexapeptides tested ($r = 0.75$, $p < 0.0001$; Table 3).

3. Discussion and conclusion

There are many nonlinear intramolecular interactions in the protease enzymes and in their substrates, as well as intermolecular interactions of the enzymes and substrates. This complexity makes it difficult to comprehensively understand all features governing the interactions of proteases with their substrates, in particular when we consider the very large mutational capabilities of retroviral proteases. Understanding all of these particulars and using them to design inhibitors with high inhibitory activity across multiple proteases accordingly is a daunting task. In fact, it appears technically untenable to use any of the current, commonly applied approaches in drug discovery to cover all the mutations known for targets of HIV.

The approach presented here, however, provides a general strategy. By combining a large amount of genetic and biochemical data for multiple retroviral proteases, we derived a unified proteochemometric model that simultaneously encompasses the ability of retroviral proteases and their mutants from many different species to interact with different substrates. This general approach for inclusion of multiviral species-proteases is highly advantageous for exploration and coverage of the HIV-1 proteases, interaction space. Other retroviral proteases are structurally and functionally very similar to the HIV-1 proteases, and the same amino acid residues in HIV-1 protease mutants that contribute to drug resistance can frequently be found in equivalent positions in other retroviral proteases. Indeed, the model constructed using only 'wild-type' HIV-1 protease and retroviral proteases from eight other viral species could cover the interaction space of HIV-1 protease mutants reasonably well to produce accurate predictions for many HIV-1 protease mutants; this ability is completely lacking in the structure–activity models constructed for an individual protein target.

The GKM constructed here markedly outperformed the previously reported in Ref. [19] cleavage rate model for retroviral proteases. It resulted in a noticeably better correlation of the physicochemical properties of retroviral proteases and substrates to the Michaelis constant log(K_m) compared to their correlation with the catalytic rate log (k_{cat}/K_m). The statistical validity of GKM using both new retroviral proteases and external data sets comprising new substrates was also better. According to Eq. 1 (see Section 4), the Michaelis constant K_m is influenced by the physicochemical properties of a substrate, a protease, experimental conditions, and their various interaction terms. However, substrate ordinary descriptors have the most significant influence on K_m because their regression coefficients are the largest according to the GKM model. Therefore, the physicochemical properties of the substrate amino acids have a major impact on the K_m constant, not including the fact that a particular protease may prefer a particular substrate (see explanation under Eq. 2 in Section 4). This generalization of the principal physicochemical properties of substrate amino acids allows selection of appropriate natural and artificial amino acids bearing suitable z-scales, based on identified PLS coefficients; these can then be used to construct libraries of peptides with potential low K_m values with many proteases. However, a cumulative effect of all other, less-important descriptors and interaction terms used in the GKM model may also considerably affect K_m value (see Eq. 1 in Section 4). Virtual screening of the peptide libraries to multiple wild-type and drug-resistant mutant proteases using GKM offers a simple solution to this problem; it captures the overall effect of the descriptors and interaction terms and accurately estimates substrate K_m to various retroviral proteases included in the screening process. As we demonstrate here, our approach was highly successful and resulted in peptides with inhibitory activity on the tested HIV-1 proteases. Some of these peptides did, indeed, also show highly uniform inhibitory activity against drug-resistant forms as well as a drug-sensitive form of the HIV-1 protease.

The approach presented here could, in principle, be continued. Based on the new data obtained from tests of the newly synthesized inhibitors, new proteochemometric models might be created. Based on these new models, further inhibitors with an improved ability to inhibit broadly across many resistance mutations could be designed. This type of process has the potential to be iterated

many times using any number of drug-sensitive and drug-resistant proteases until identification of a compound with the desired properties. A compound that could broadly inhibit many mutated forms of a protease might more broadly arrest genetically fit proteases and thus reduce the targets for the evolution of a virus into resistant strains. Therefore, such a compound might reduce the problem of resistance development in HIV.

Proteochemometrics has no intrinsic limitations in terms of how many proteases can be analyzed at the same time. This feature places it in a special position compared to other contemporary methods in drug design, which are generally directed at one target at a time. The proteochemometrics approach should have wide-ranging implications in facilitating studies of interaction space and the resistance mechanisms of HIV, as well as of other pathogens exhibiting resistance to drugs because of target protein heterogeneity.

4. Experimental

4.1. Data

The data set, collected in a survey covering publicly available data from 17 years of retroviral protease research, is available in Supplementary Table 1 and includes K_m data for 9 different retroviruses, that is, 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). The data set included a total of 654 Michaelis constants (K_m) observations for retroviral protease substrates. In some cases the K_m constant was reported as an inequality, for example: 'K_m < 0.01 mM' or 'K_m > 10 mM'. In these cases we assumed the K_m to be equal to the value shown. Some of the data was generated in house (Supplementary Table 2).

4.2. Data pre-processing

4.2.1. Description of proteases

The 63 'wild-type' and mutated retroviral protease sequences included in the study (Supplementary Table 3) were structurally aligned as reported earlier.[19] According to this alignment, 94 amino acids could be fully aligned over all proteases and without using any 'gaps' (9 conserved and 85 non-conserved positions). We selected the non-conserved 85 amino acid positions and described each of them by their five principal physicochemical properties 'z-scales'.[20] These z-scales are orthogonal to each other and represent roughly hydrophobicity (z_1), steric properties (z_2), polarizability (z_3), polarity and electronic effects of amino acids (z_4, z_5). Accordingly, 85 protease positions were described by 85 × 5=425 descriptors (also termed ordinary protease descriptors) which thus comprised the physicochemical property space information of the series of proteases used herein.

4.2.2. Description of substrates

The length of retroviral substrates was restricted to octapeptides (P_4–P_3–P_2–P_1–P'_1–P'_2–P'_3–P'_4, where P_4 denotes substrate N-terminus amino acid and P'_4—C-terminus substrate amino acid; the scissile bond being located between the P_1 and P'_1 amino acids). This approach was elected because eight substrate amino acid residues are considered to interact with eight corresponding retroviral protease subsites (S_4–S_3–S_2–S_1–S'_1–S'_2–S'_3–S'_4). We described each of the eight amino acids of the substrates by the same five z-scales as above, which thus yielded in total 8 × 5 = 40 descriptors per substrate. These descriptors (herein also termed ordinary substrate descriptors) thus encompassed the physicochemical space information of the series of substrates used herein.

4.2.3. Description of assay conditions and kinetics of experimental data

In addition to protease and substrate descriptors, we also included eight descriptors for the assays, that is, pH, sodium chloride, 2-mercaptoethanol, EDTA, DMSO, dithiothreitol, nonidet-P40, and glycerol concentrations. Michaelis constants (K_m) were converted to μM units for all experiments, followed by their decimal logarithmic transformation, log (K_m).

4.2.4. Description of inter-dependences of proteases, substrates, and assays

Interaction term descriptor blocks were constructed by multiplying separately each protease descriptor with each other protease descriptor (protease × protease interaction term descriptors block), each substrate descriptor with each protease descriptor (substrate × protease interaction term descriptor block), each substrate descriptor with each other substrate descriptor (substrate × substrate interaction term descriptors block), each substrate descriptor with each assay descriptor (substrate × assay interaction term descriptor block), each assay descriptor with each protease descriptor (assay × protease interaction term descriptor block) and each assay descriptor with each other assay descriptor (assay × assay interaction term descriptor block). These interaction term descriptor blocks were used in the modeling in various combinations with the blocks formed from, respectively, ordinary protease, substrate, and assay descriptors to assess their significance and find the best combination of the descriptor blocks required to achieve an optimal model. All ordinary protease, substrate, and assay block descriptors were mean-centered and scaled to unit variance prior to computation of interaction terms. In addition we applied block-scaling for each type of descriptor block in order to account for differences in number and mutual correlation of descriptors in each block.[33] (Using block-scaling avoids situations where large blocks of descriptors mask small ones).

4.3. Multivariate modeling and data analysis

All observations listed in Supplementary Table 1 were used for the construction of generalized retroviral protease K_m models, using ordinary protease, substrate, and assay descriptor blocks as well as interaction term descriptor blocks in various combinations, as detailed in Table 1. Models STM1–STM6 used only 200 experiments for the HXB2 HIV-1 protease and were termed 'single-target models' (STM) (Table 1). The preprocessed descriptors were correlated to log (K_m) by partial least squares (PLS) regression using Simca-P+ 11.5 software (http://www.umetrics.com) and validated as described below. Models were considered acceptable if $R^2 > 0.7$ and $Q^2 > 0.4$.[33,34] We also considered iR^2 and iQ^2 parameters, which should not exceed 0.4 and 0.05, respectively, for a valid, not overfitted, PLS model.[35]

4.4. Validation of the models

The goodness-of-fit of PLS regressions were measured by the unit-less parameter R^2, which can range between 0.0 and 1.0. R^2 indicates the fraction of the sum of squares explained by the model, and a higher R^2 values signifies that the model fits the data better.[33,34] We also used the root mean square error of estimation (RMSEE) to calculate the internal error within the model.

Cross-validation (CV) is a method for evaluation of the mean generalization accuracy for a regression model on sets of data. In CV the data set is randomly divided into k parts (10-fold CV was used herein) and each one of these parts is then used to test a model fitted to the remaining $k - 1$ parts. This results in a cross-validated regression coefficient, Q^2, where a higher Q^2 denotes better predictability.[34,36,37] The 10-fold CV was repeated 20 times on

different data set subdivisions in order to estimate the variability of the mean prediction accuracy (Q^2_{mean}) for every model.

For permutation validation the dependent variable log (K_m) was repeatedly and randomly permutated, yielding new data set samples with replacements from the original data set (100 randomly permutated data sets were used for each model validation).[15,38] New models were then built on the permutated data sets and R^2, Q^2 and correlation coefficients between original and permutated response values were estimated. Finally, intercept values for R^2 (iR^2) and Q^2 (iQ^2) reflecting R^2 and Q^2 of the models constructed on the randomly permutated data sets were computed.[15] A small iR^2 value ($iR^2 < 0.4$) signifies that there is little chance-correlation in the original model, whereas a negative iQ^2 indicates that it is impossible to get predictive models based on randomly permutated data.[35]

'Leave-one-protease-out' validation of GKM was performed by entirely excluding all data for the proteases of one retroviral strain and then predicting the excluded data using the model constructed from the remaining data. Because of the small number of observations available for the BLV, FIV, and EIAV proteases, 'leave-one-protease-out' validation was not considered feasible for these. In the case of HIV-1 proteases, the data for HXB2 HIV-1 protease and HIV-1 proteases with five artificial stabilizing mutations (Q7K + L33I + L63I + C67A + C95A) were kept in the model, and the external predictions were performed for the remaining 23 HIV-1 proteases with drug-resistant mutations.

GKM was also experimentally validated on an independent validation set consisting of the K_m data for the HXB2 HIV-1 protease, its three mutants with drug-resistance mutations (I84V, L90M, and I84V + L90M), and 15 substrates of diverse structures (Supplementary Table 2, numbers 4–18). We used the RMSEP to evaluate the model's ability to predict the external data.[33]

4.5. Analysis of the generalized K_m model

For the GKM model, the regression equation can be expressed as follows:

$$\log(K_m) = \overline{\log(K_m)} + \sum_i^N k_i x_i^{sub} + \sum_j^M k_j x_j^{PR} + \sum_f^L k_f x_f^{assay} + \sum_i^N x_i^{sub}\left(\sum_j^M k_{ij} x_j^{PR}\right) + \sum_i^N x_i^{sub}\left(\sum_f^L k_{if} x_f^{assay}\right) + \sum_{a,b,a<b}^{0.5\times N(N-1)} k_{ab} x_a^{sub} x_b^{sub} \quad (1)$$

where $\overline{\log(K_m)}$ is the average logarithmic K_m; N, M, and L the number of descriptors in substrate, protease, and assay blocks, respectively; k_i, k_j, k_f, k_{ij}, k_{if}, and k_{ab} the regression coefficients for substrate descriptors, protease descriptors, assay descriptors, substrate × protease interaction terms, substrate × assay interaction terms, and substrate × substrate interaction terms, respectively; and x^{sub}, x^{PR}, and x^{assay} substrate, protease, and assay ordinary descriptors.

The difference in activities between some proteases PR_1 and PR_2 for some substrate Z can be expressed as follows:

$$\Delta\log(K_m)_{PR1-PR2-Z} = \log(K_m)_{PR1-Z} - \log(K_m)_{PR2-Z} = \sum_j^M k_j(x_j^{PR1} - x_j^{PR2}) + \sum_i^N x_i^Z\left(\sum_j^M k_{ij}(x_j^{PR1} - x_j^{PR2})\right) \quad (2)$$

According to Eq. 2, the difference in K_m between any two proteases PR_1 and PR_2 for some particular substrate Z depends on descriptor differences in proteases PR_1 and PR_2 (i.e., $x_j^{PR1} - x_j^{PR2}$), the size of PLS coefficients for the ordinary protease descriptors (k_j), and protease × substrate interaction term descriptors (k_{ij}); that is, the larger the PLS coefficient, the larger the influence a corresponding descriptor has on K_m. Analysis of GKM demonstrated that substrate, assay ordinary descriptor blocks, and a substrate × substrate interaction term descriptor block have the largest PLS coefficients in the model, whereas a protease ordinary descriptor block and protease × substrate interaction terms have the smallest. Therefore, if some substrate Z is designed so that its descriptors (z-scales) are large and have opposite tokens to corresponding PLS coefficients, as shown in Figure 3, then substrate Z should demonstrate uniformly low K_m against many retroviral proteases because descriptors of retroviral proteases have smaller coefficients and accordingly a lower impact on K_m in general.

All descriptors used for building the GKM were normalized by mean-centering and scaling to unit variance, allowing us to compare their coefficients. The largest coefficients indicate the most important descriptors for the GKM model's outcome and were in some cases used directly (Fig. 3). We also identified descriptors in the substrate × substrate interaction term block with the largest absolute values of their coefficients. The five largest substrate × substrate interaction terms are the ones shown in Table 2. In a similar way, we analyzed the coefficients for the assay descriptor block to find the assay constituents that most influenced the Michaelis constant K_m (Table 2).

To localize the most important retroviral protease amino acids (shown in Table 2), we compared the sums of absolute values of the 5 z-scale descriptor coefficients for each of the 85 aligned protease amino acids. This procedure allows capture of combined physicochemical property effects induced by each of the amino acids considered. The 8 amino acid positions having the largest sum of their 5 z-scale coefficients and consequently having large impacts on K_m are shown in Table 2.

We further analyzed the GKM coefficients for the substrate protease descriptor block. Because every substrate and protease amino acid was described by 5 z-scale descriptors, every substrate–protease amino acid pair creates 25 interaction terms (5 × 5). To identify the most important substrate–protease interactions, we computed the absolute values sum of the coefficients for each protease–substrate amino acid pair and then compared the sums for each pair. As a result, the 10 most important substrate–protease inter-dependences were identified (Table 2).

4.6. Hexapeptide library design and in silico screening

The library of hexapeptides was based on a set of amino acids suitable for each substrate position P_3–P'_3, indicated by the PLS coefficients of the ordinary descriptors of GKM (Fig. 3). We also included artificial amino acids to explore the physicochemical space of substrates as much as possible, and selected amino acids were as follows: for the P_3 position, they were Hph, Nal, or Cph (for abbreviation of artificial amino acids see below); for P_2, Val, Nva, Nle, or Cha; for P_1, Lys, Met, Dap, or Dab; for P'_1, Arg, Glu, Har, or Aph; for P'_2, Phe, Tyr, Btr, Nal, Hph, Phg, or Bph; and for the P'_3 position, Leu, Nle, Cha, or Nva. The initial constructed virtual library consisted of $3 \times 4 \times 4 \times 4 \times 7 \times 4 = 5376$ hexapeptides. In order to identify non-cleavable sequences for the HXB2 HIV-1 protease we used the cleavability model described in Ref. [19] to screen the library. Since the cleavability model considers octapeptides we placed Ala at both the P_4 and P'_4 positions to all hexapeptides during the virtual screening. This resulted in 4469 non-cleavable sequences. We then used GKM to further virtually screen the library predicted as non-cleavable and selected peptides having a predicted

K_m < 3 μM for HXB2 HIV-1 protease and its 3 mutants L90 M, I84 V and I84V + L90M (4154 sequences). (Inhibition assay conditions used for in silico hexapeptide library screening by the GKM were pH 5.0 and sodium chloride 1.1 M concentration. All other assay descriptors were set to 0). From this smaller library we randomly selected 10 hexapeptides for experimental evaluation, allowing at most three amino acids to be identical at the corresponding positions between any two peptides in the chosen test set (Table 3, numbers 1–10).

A set of diverse hexapeptides (Table 3, numbers 11–20) was available in our lab from various earlier projects (unpublished data). All of these hexapeptides were predicted by the cleavability model[19] to be uncleavable and were predicted by GKM to have very large K_m values (Table 3). This set of diverse-structure hexapeptides was used as a negative control test for the GKM validation.

4.7. Statistical tests

The Q^2 values calculated for 10-fold CVs did not follow a normal distribution, which was revealed from the samples of 20 repeats. In order to assess the difference in Q^2 between two models we therefore used the nonparametric Wilcoxon signed-rank test.[39]

The Pearson correlation coefficient (r) values for the observed versus GKM predicted log (K_m) was determined and the statistical significance, p, of the correlation assessed (Figs. 1 and 2). The p-value obtained is the probability that a correlation (in the positive direction) would be seen by chance if there was no real linear relationship between observed and predicted log (K_m) values. The significance tests were one-sided. The test of correlation and all significance tests were performed by an in-house add-in to the Excel program (Microsoft).

4.8. Synthesis of peptides

Hexapeptides numbers 1–20 of Table 3 were synthesized by solid-phase peptide synthesis using an automated multiple peptide synthesizer (MultiPep; Intavis AG Bioanalitical Instruments, Germany, http://www.intavis.com). Reagents were purchased from Fluka (http://www.fluka.org), Applied Biosystem (http://www.appliedbiosystems.org), Bachem (http://www.bachem.com), or Novabiochem (http://www.emdbiosciences.com/html/NBC/home.html). The following amino acid derivatives were used in the synthesis: Fmoc-Ala-OH,[†] Fmoc-Arg(Pbf)-OH, Fmoc-Asn(Trt)-OH, Fmoc-Asp(O*t*-Bu)-OH, Fmoc-Cys(Trt)-OH, Fmoc-Gln(Trt)-OH, Fmoc-Glu(O*t*-Bu)-OH, Fmoc-Gly-OH, Fmoc-His(Trt)-OH, Fmoc-Ile-OH, Fmoc-Leu-OH, Fmoc-Lys(Boc)-OH, Fmoc-Met-OH, Fmoc-Phe-OH, Fmoc-Pro-OH, Fmoc-Ser(*t*-Bu)-OH, Fmoc-Thr(*t*-Bu)-OH, Fmoc-Trp(Boc)-OH, Fmoc-Tyr(*t*-Bu)-OH, Fmoc-Val-OH, Fmoc-Aph(Boc)-OH, Fmoc-Bph-OH, Fmoc-Btr-OH, Fmoc-Cha-OH, Fmoc-Cph-OH, Fmoc-Dab(Boc)-OH, Fmoc-Dap(Boc)-OH, Fmoc-Har(Pmc)-OH, Fmoc-Hph-OH, Fmoc-Nal-OH, Fmoc-Nle-OH, Fmoc-Nva-OH and Fmoc-Phg-OH. PyBOP was used as an activating reagent and Tenta Gel amide resin (capacity 0.26 mmol/g) as a polymeric support.

The peptides were synthesized in 5 μmol scale using the automated standard protocol optimized for Fmoc chemistry provided with the MultiPep synthesizer. Each cycle included deprotection of Fmoc group by 20% piperidine in DMF and washing of the support with DMF; coupling (i.e., the N-deblocked peptidyl-resin was treated with the solution of Fmoc amino acid derivative, PyBOP and NMM in DMF for 25 min) and washing of the support with DMF; capping (i.e., treatment of the polymer with the 2% solution of acetic anhydride in DMF for 5 min) and washing of the support with DMF. The final synthetic step on MultiPep included deprotection by 20% piperidine in DMF (peptides numbers 1–10, Table 3) or treatment of the polymer with the 2% solution of acetic anhydride in DMF (peptides numbers 11–20, Table 3), washing of the support with DMF and CH_2Cl_2 and drying. The peptide was deprotected and cleaved from the resin with deprotection mixture (TFA–triisopropylsilane-1,2-ethanedithiol–water, 92.5:2.5:2.5:2.5) for 3 h at room temperature, triturated with *tert*-butyl-methyl ether, taken up in MeCN/water, lyophilized, purified by HPLC and their structures were confirmed by mass spectrometry. Analytical HPLC was performed on a Waters (http://www.waters.com) system (Millenium32 workstation, 2690 Separation Module, 996 photodiode array detector) equipped with Vydac RP C18 90 Å reversed-phase column (2.1 × 250 mm; http://www.vydac.com).

Small-scale preparative HPLC was carried out on a system consisting of a 2150 HPLC Pump, 2152 LC Controller and 2151 variable wavelength monitor (LKB, Sweden) and Vydac RP C_{18} column (10 mm 250 mm, 90 Å, 201HS1010), with the eluent, being an appropriate concentration of MeCN in water + 0.1% TFA, a flow rate 5 mL/min, and detection at 280 nm. Freeze-drying was carried out at 0.01 bar on a Lyovac GT2 Freeze-Dryer (Steric Finn-Aqua; http://www.steric.com) equipped with a Trivac D4B (Leybold Vacuum; http://www.oerlikon.com) vacuum pump and a liquid nitrogen trap.

Peptides were checked by LC/MS using a Perkin Elmer PE SCIEX API 150EX instrument equipped with a turboionspray ion source (PerkinElmer Life And Analytical Sciences; http://las.perkinelmer.com) and a Dr. Maisch Reprosil-Pur C18-AQ HPLC column (5 μm, 150 mm × 3 mm; http://wwwdr-maish.com), using a gradient formed from water and acetonitrile with 5 mM ammonium acetate additive.

When not otherwise specified chemicals were of reagent grade from Sigma (http://www.sigmaaldrich.com).

4.9. Enzyme assay of substrates and hexapeptide inhibitors

The HXB2 clone of the HIV-1 protease and its three mutant (I84V, L90M, and I84V + L90M) clones were available in our laboratory and were kind gifts of Prof. Helena Danielson, Uppsala University[40]. All assays were performed in black 96-well plates (Nunc) using a PolarstarOptima microplate reader (excitation and emission waves were 355 nm and 490 nm, respectively). The reaction buffer contained 0.1 M acetic acid and 1.1 M sodium chloride (pH 5.0 was achieved by titration with a sodium hydroxide solution). Substrate and inhibitor stock solutions were 1 mM and dissolved in DMSO/water (1:2). A typical inhibition reaction mixture contained 2.25 μg of the substrate (substrate Nr.1 in Supplementary Table 2), variable concentrations of hexapeptides and 35 ng of HXB2 HIV-1 protease or 70 ng of HIV-mutant enzymes (total reaction volume was 100 μL). HIV-1 protease was incubated 10 min at 37 °C with the hexapeptide inhibitor (total solution volume 70 μL). 30 μL of substrate solution were added after the incubation and reaction continued for further 30 min at 37 °C (cycle time 60 s, 5 s shaking after each cycle). Inhibitor dilution series started at 100 μM of hexapeptide. Each next dilution comprised a factor of two of the previous dilution and 11 dilutions were used for the inhibition assay of each hexapeptide. A control test (incubation of HIV-1 protease with the substrate but without a respective

[†] *Abbreviations used*: Boc, *tert*-butoxycarbonyl; *t*-Bu, *tert*-butyl; DMF, *N*,*N*-dimethylformamide; Fmoc, fluoren-9-yl-methoxycarbonyl; MeCN, acetonitrile; NMM, *N*-methylmorpholine; Pbf, 2,2,4,6,7-pentamethyldihydrobenzofuran-5-sulfonyl; Pmc, 2,2,5,7,8-pentamethylchromane-6-sulfonyl; PyBOP, benzotriazole-1-yl-oxy-tris-pyrrolidino-phosphonium hexafluorophosphate; TFA, trifluoroacetic acid; Hph, homophenylalanine; Nal, 3-(2-naphthyl)alanine; Cph; *p*-cholophenylalanine; Nva, norvaline; Cha, cyclohexylalanine; Nle, norleucine; Dab, 2,4-diaminobutyric acid; Dap, 2,3-diaminopropionic acid; Har, homoarginine; Aph, *p*-aminophenylalanine; Phg, phenylglycine; Btr, *O*-benzyltyrosine; Bph, *p*-bromophenylalanine. All amino acids used for synthesis are L-isomers.

hexapeptide inhibitor) was also performed. In total each inhibition assay experiment comprised of 12 data points.

The kinetic data was analyzed by non-linear fit using GRAFIT program and the basic equation for Michaelis-Menten kinetics.[41] The obtained Michaelis constants K_m were converted into μM units before use in the data analyses. The inhibition assay data was fitted by non-linear regression analysis using GRAFIT program[41] and obtained IC_{50} values were converted into K_i values according to the equation of Cheng and Prusoff.[42] Each experiment was repeated at least three times, and the average value was taken as a final result. A K_i value of 200 μM was set for hexapeptides marked as 'NI' in Table 3 for the purpose of testing the log (K_i)/log (K_m) correlations.

Acknowledgements

We are indebted to Dr. Helena Danielson for the generous gift of the clones used in expression of the HXB2 and mutant HIV-1 proteases. Financial support was provided by Swedish International Development Cooperation Agency (HIV-2006-019) and the Swedish Research Council (04X-05957).

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.bmc.2009.05.045.

References and notes

1. Erickson, J. W.; Stanley, K. B. *Annu. Rev. Pharmacol. Toxicol.* **1996**, *36*, 545.
2. Abdel-Rahman, H. M.; Kimura, T.; Kiso, A.; Nezami, A.; Freire, E.; Hayashi, Y.; Kiso, Y. *Biol. Chem.* **2004**, *385*, 1035.
3. Li, M.; Laco, G. S.; Jaskolski, M.; Rozycki, J.; Alexandratos, J.; Wlodawer, A.; Gustchina, A. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102*, 18332.
4. Ohtaka, H.; Muzammil, S.; Schon, A.; Velazquez-Campoy, A.; Vega, S.; Freire, E. *Int. J. Biochem. Cell Biol.* **2004**, *36*, 1787.
5. Ohtaka, H.; Freire, E. *Prog. Biophys. Mol. Biol.* **2005**, *88*, 193.
6. Wlodawer, A.; Gustchina, A. *Biochim. Biophys. Acta* **2000**, *1477*, 16.
7. Dunn, B. M.; Goodenow, M. M.; Gustchina, A.; Wlodawer, A. *Genome Biol.* **2002**, *3*, 3006. 1.
8. Freire, E. *Nat. Biotechnol.* **2002**, *20*, 15.
9. Randolph, J. T.; DeGoey, D. A. *Curr. Top. Med. Chem.* **2004**, *4*, 1079.
10. Dash, C.; Kulkarni, A.; Dunn, B.; Rao, M. *Crit. Rev. Biochem. Mol. Biol.* **2003**, *38*, 89.
11. Beck, Z. Q.; Morris, G. M.; Elder, J. H. *Curr. Drug Targets Infect. Disord.* **2002**, *2*, 37.
12. Prusis, P.; Uhlen, S.; Petrovska, R.; Lapinsh, M.; Wikberg, J. E. S. *BMC Bioinform.* **2006**, *7*, 167.
13. Lapinsh, M.; Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. *Mol. Pharmacol.* **2002**, *61*, 1465.
14. Lapinsh, M.; Prusis, P.; Mutule, I.; Mutulis, F.; Wikberg, J. E. S. *J. Med. Chem.* **2003**, *46*, 2572.
15. Prusis, P.; Lundstedt, T.; Wikberg, J. E. S. *Protein Eng.* **2002**, *15*, 305–311.
16. Wikberg, J. E.; Mutulis, F.; Mutule, I.; Veiksina, S.; Lapinsh, M.; Petrovska, R.; Prusis, P. *Ann. NY Acad. Sci.* **2003**, *994*, 21.
17. Wikberg, J. E. S.; 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.
18. Kontijevskis, A.; Petrovska, R.; Mutule, I.; Uhlen, S.; Komorowski, J.; Prusis, P.; Wikberg, J. E. S. *Proteins* **2007**, *69*, 83.
19. Kontijevskis, A.; Prusis, P.; Petrovska, R.; Yahorava, S.; Mutulis, F.; Mutule, I.; Komorowski, J.; Wikberg, J. E. S. *PLoS Comput. Biol.* **2007**, *3*, e48.
20. Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. *J. Med. Chem.* **1998**, *41*, 2481.
21. Gustchina, A.; Weber, I. T. *FEBS Lett.* **1990**, *269*, 269.
22. Szeltner, Z.; Polgar, L. *J. Biol. Chem.* **1996**, *271*, 32180.
23. Debnath, A. K. *Curr. Pharm. Des.* **2005**, *11*, 3091.
24. Kurup, A.; Mekapati, S. B.; Garg, R.; Hansch, C. *Curr. Med. Chem.* **2003**, *10*, 1679.
25. Johnson, V. A.; Brun-Vezinet, F.; Clotet, B.; Kuritzkes, D. R.; Pillay, D.; Schapiro, J. M.; Richman, D. D. *Top HIV Med.* **2006**, *14*, 125.
26. 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. *J. Virol.* **1994**, *68*, 2016.
27. Tozser, J.; Weber, I. T.; Gustchina, A.; Blaha, I.; Copeland, T. D.; Louis, J. M.; Oroszlan, S. *Biochemistry* **1992**, *31*, 4793.
28. Hyland, L. J.; Tomaszek, T. A., Jr.; Meek, T. D. *Biochemistry* **1991**, *30*, 8454.
29. Ido, E.; Han, H. P.; Kezdy, F. J.; Tang, J. *J. Biol. Chem.* **1991**, *266*, 24359.
30. Polgar, L.; Szeltner, Z.; Boros, I. *Biochemistry* **1994**, *33*, 9351.
31. Beck, Z. Q.; Hervio, L.; Dawson, P. E.; Elder, J. H.; Madison, E. L. *Virology* **2000**, *274*, 391.
32. Jordan, S. P.; Zugay, J.; Darke, P. L.; Kuo, L. C. *J. Biol. Chem.* **1992**, *267*, 20028.
33. Simca-P+ 10.0 *Manual*, Umetrics AB: Umeå, Sweden, 2002.
34. Lundstedt, T.; Seifert, E.; Abramo, L.; Thelin, B.; Nystrom, A.; Pettersen, J.; Bergman, R. *Chemometr. Intell. Lab. Syst.* **1998**, *42*, 3.
35. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Trygg, J.; Wikström, C.; Wold, S. *Multi- and Megavariate Data Analysis*, 2nd ed.; Umetrics AB: Umeå, Sweden, 2006.
36. Wold, S. *Technometrics* **1978**, *20*, 397.
37. Wakeling, I. N.; Morris, J. J. *J. Chemometr.* **1993**, *7*, 291.
38. Efron, B. *J. Am. Stat. Assoc.* **1987**, *82*, 171.
39. Wilcoxon, F. *Biometrics* **1945**, *1*, 80.
40. Danielson, H.; Lindgren, M. T.; Markgren, P. O.; Nillroth, U. *Adv. Exp. Med. Biol.* **1998**, *436*, 99.
41. Bardsley, WG.; McGinlay, P. B. *J. Theor. Biol.* **1989**, *139*, 85.
42. Cheng, Y.; Prusoff, W. H. *Biochem. Pharmacol.* **1973**, *22*, 3099.