Bioorganic & Medicinal Chemistry 16 (2008) 9369–9377

Contents lists available at ScienceDirect

Bioorganic & Medicinal Chemistry

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

Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases

Peteris Prusis [a,†], Maris Lapins [a,†], Sviatlana Yahorava [a], Ramona Petrovska [a], Pornwaratt Niyomrattanakit [b], Gerd Katzenmeier [b], Jarl E. S. Wikberg [a,*]

[a] Department of Pharmaceutical Biosciences, Uppsala University, Box 591 BMC, SE751 24 Uppsala, Sweden
[b] Laboratory of Molecular Virology, Institute of Molecular Biology and Genetics, Mahidol University, Salaya Campus, Phutthamonthon 4 Road, Nakornpathom 73170, Thailand

ARTICLE INFO

Article history:
Received 15 March 2008
Revised 7 August 2008
Accepted 20 August 2008
Available online 13 September 2008

Keywords:
Dengue proteases
Proteochemometrics
Substrate library
Peptide library
Library design
Statistical molecular design
Molecular recognition modeling

ABSTRACT

The prime side specificity of dengue protease substrates was investigated by use of proteochemometrics, a technology for drug target interaction analysis. A set of 48 internally quenched peptides were designed using statistical molecular design (SMD) and assayed with proteases of four subtypes of dengue virus (DEN-1–4) for Michaelis (K_m) and cleavage rate constants (k_{cat}). The data were subjected to proteochemometrics modeling, concomitantly modeling all peptides on all the four dengue proteases, which yielded highly predictive models for both activities. Detailed analysis of the models then showed that considerably differing physico-chemical properties of amino acids contribute independently to the K_m and k_{cat} activities. For k_{cat}, only P1′ and P2′ prime side residues were important, while for K_m all four prime side residues, P1′–P4′, were important. The models could be used to identify amino acids for each P′ substrate position that are favorable for, respectively, high substrate affinity and cleavage rate.

1. Introduction

Dengue fever has been known for more than two hundred years. The disease is caused by the dengue virus, a member of *Flaviviridae* family, which is transmitted to humans by mosquitoes of the species *Stegomiya aegypti (formerly Aedes)*. There are four closely related but antigenically distinct dengue virus serotypes (DEN-1–4) for which immunity to one serotype does not protect against infection by another.[1,2] Infections occur primarily in the tropics, where the virus threatens a large portion of the population. Its global distribution is comparable to that of malaria, and an estimated 2.5 billion people live in areas at risk for epidemic transmission.[3–5]

Dengue causes a spectrum of clinical symptoms ranging from mild, uncomplicated dengue fever to the severe forms of dengue hemorrhagic fever and dengue shock syndrome.[4–7] Dengue fever can cause aches, pains, headaches, and high fever, and is sometimes called "breakbone fever" because of the pain associated with it. Dengue hemorrhagic fever and shock syndrome are severe forms with hemorrhaging that may cause dramatic loss of blood pressure. Nearly 5% of the ~1 million dengue hemorrhagic fever cases occurring each year are fatal.[4]

Accordingly, there is considerable interest in developing therapeutics against dengue. Vaccine candidates for all four serotypes derived from live attenuated or chimeric viruses are in clinical trial.[8] However, there is presently neither useful vaccine nor antiviral drug available.

The genomes of the dengue viruses consist of an 11-kb single positive-stranded RNA that encodes 3 structural (C, prM, and E) and seven non-structural proteins (NS1, NS2A, NS2B, NS3, NS4A, NS4B, and NS5).[9] The correct processing of these proteins is essential for virus replication and requires host proteases such as signalase and furin[10] and a two-component viral protease, NS2B/NS3.[9,11]

The dengue virus NS3 protease is an attractive target for development of therapeutic inhibitors of dengue. The enzyme is vital for the post-translational proteolytic processing of the dengue polyprotein precursor and is essential for viral replication and maturation of infectious virons.[11–13] The NS3 protease catalyzes the cleavage of the viral polyprotein precursor in the non-structural region, in *cis* at the NS2A/NS2B and NS2B/NS3 junctions and in *trans* at the NS3/NS4A and NS4B/NS5 sites[11,14–16] (as well as at additional sites within the viral capsid protein, NS2A, NS4A, and within NS3 itself).[17–20] A trypsin-like protease domain with a classical

Abbreviations: SMD, statistical molecular design; K_m, Michaelis constant; k_{cat}, conversion rate constant; DEN-1–4, dengue virus serotypes 1–4; Abz, *o*-aminobenzoic acid; nY, 3-nitrotyrosine; PCA, principal component analysis; PLS, partial least-squares projections to latent structures.

* Corresponding author. Tel.: +46 18 471 4238; fax: +46 18 55 9718.
E-mail address: Jarl.Wikberg@farmbio.uu.se (J.E.S. Wikberg).

† These authors contributed equally to this work.

0968-0896/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmc.2008.08.081

serine protease catalytic triad (His51, Asp75, and Ser135) was originally identified in the N-terminal region of the 69 kDa NS3 protein[15] and the minimum sequence which supports protease activity was mapped to 167 residues of NS3.[21] The protease activity is enhanced by the NS2B protein (at least 40 amino acids of it are required), which acts as cofactor for the NS3 protease.[14,22,23]

The NS3 protease-cleavage sites in the viral polyprotein have pairs of dibasic amino acids (i.e., RR, RK, KR, or, at times, QR) at the P2 and P1 positions, and small non-branched amino acid at the P1′ position.[9,12,24–27] Our earlier study suggested that basic amino acids at the P3 and P4 positions improved substrate cleavage[27], and another study by Li et al.[26] recently suggested that aliphatic amino acids may be favorable at the P4 position. Our previous work showed that removal of the methionine at the P4′ position in a peptide which spanned both P4–P1 and P1′–P4′ of the capsid protein cleavage site in DEN-2, namely Abz–RRRR↓SAGM–nY–NH_2 peptide (i.e., resulting in Abz–RRRR↓SAG–nY–NH_2), led to improved k_{cat} activity.[27] Li and colleagues investigated the specificity of the P′ side by comparing fluorescence-detected initial velocities for pools of peptide substrates, holding one amino acid fixed while varying the other three positions, which resulted in mixtures of 19 amino acids at each position.[26] The results of this study, which included all four viral subtypes of the protease, showed that serine was the most preferred amino acid in the P1′ and P3′ positions of the substrate, whereas the P2′ and P4′ positions tolerated a broad diversity of amino acids. Two earlier studies were also dedicated to the design of dengue protease inhibitors, essentially based on peptidomimetics mirroring the dibasic P1–P2 site of dengue protease substrates.[28,29]

We have developed a novel approach to study ligand–protein interactions, proteochemometrics.[30,31] Proteochemometrics analyzes experimentally determined interaction data for series of ligands with respect to series of proteins, correlating interaction data to the physico-chemical and/or structural descriptions of both the ligands and proteins. Proteochemometric models are thereby created, which reveal the properties of both interaction partners that determine their activities and specificities. Proteochemometrics has been successfully applied to various classes of G-protein coupled receptors,[32–34] antibodies[35] and aspartate proteases.[36] The aim of the present study was to extend our knowledge on the requirements for dengue virus NS3 protease substrates. To this end, we investigated the P′ side specificity of substrates for the NS3 protease from all four serotypes using a large set of internally quenched peptides, separately analyzing the rate of substrate cleavage (k_{cat}) and substrate affinity (K_m) by proteochemometric modeling.

2. Results and discussion

2.1. Design of substrate library

To investigate the roles of the P1′–P4′ positions of dengue protease substrates, we designed a library of internally quenched peptides having the general structures Abz–RRRR↓XXXX–nY–NH_2 (Abz, *o*-aminobenzoic acid; nY, 3-nitrotyrosine) and Abz–RRRR↓XXX–nY–NH_2. In preliminary studies, we deduced a set of amino acids for each of the prime sequence positions that had great potential to result in active substrates for the DEN-2 protease. These amino acids were identified from tests of a series of 32 octapeptides with amino acids selected from the P4–P1 and P1′–P4′ positions of natural DEN-2 polyprotein cleavage sites. Cleavage kinetics of these peptides by the DEN-2 protease were investigated and the obtained initial reaction velocity values were subjected to quantitative structure–activity relationship (QSAR) modeling (data not shown). According to this analysis, the peptides that were most readily cleaved had four arginines in the P4–P1 positions, similarly to the capsid protein cleavage site in the DEN-2 virus. Based on the predictions of this QSAR model, we then identified sets of amino acids for the P1′–P4′ positions that should ensure cleavability of the substrates. The following sets of amino acids were identified: A, N, D, G, H, and S for the P1′ position; A, N, D, G, H, L, P, S, T, and W for the P2′ position; C, G, P, S, T, and V for the P3′ position; and N, D, C, H, L, F, or none for the P4′ position (thus giving a total of $6 \times 10 \times 6 \times 7 = 2520$ possible amino acid combinations for P1′–P4′).

An informative peptide library was created from the above sets by applying the principles of statistical molecular design,[37,38] using D-optimal design as implemented in MODDE 6.0 software (Umetrics AB, Sweden). This design technique maximizes the diversity in the set of objects comprised by the design matrix **X**. The optimality criterion in generating D-optimal design is the determinant of the $\mathbf{X}'\mathbf{X}$ matrix, which is an overall measure of the information in **X**.[39] Applying D-optimal design gave a set of 48 peptides (peptides Nr 1–48 in Table 1) that covered the maximal volume of the whole multivariate space of possible amino acid combinations. The quality of design was characterized by the following parameters: $\log(\det \mathbf{X}'\mathbf{X}) = 24.7$ and G-efficiency = 66.1.[40] Hereafter, this library of 48 peptides is referred to as the work set.

The validity of the proteochemometric models was assessed by an external test set for which we used eight previously prepared peptides having the same general structure as those above.[27] The prime side sequences of these peptides had been derived from the native cleavage sites in the dengue polyprotein, indicated as peptides Nr 49–56 shown in Table 1.

2.2. Kinetic characterization of substrate library

For the four subtypes of the dengue protease, we determined values of K_m and k_{cat} using the work set of 48 substrates (see Section 4 for peptide synthesis, expression of proteases, and determination of kinetic constants). The data obtained represented the averages from at least three independent measurements for each protease–substrate pair and are summarized in Table 1. Only one peptide failed to be cleaved by all four proteases. Moreover, the DEN-1 protease failed to cleave one additional substrate, while the DEN-3 protease failed to cleave seven substrates. Some substrates gave marked substrate inhibition, which precluded calculations of K_m and k_{cat} values. Interestingly, there was no apparent correlation between the Michaelis constant and the cleavage rate. Thus, for the DEN-1, DEN-3, and DEN-4 proteases there were no correlations whatsoever between the logarithmically transformed K_m and k_{cat} values ($r^2 < 0.05$), while for the DEN-2 protease there was a marginally negative correlation ($r^2 = 0.14$). These results thus suggest that mechanisms for substrate binding and cleavage are independent for the dengue proteases.

2.3. Results of proteochemometric modeling

As detailed in Section 4, each of the four P1′–P4′ residues of the substrates was characterized by 6 quantitative descriptors representing hydrophilicity (zz_1), size (zz_2), polarity (zz_3), charge (C7.4), rigidity (t1-Rig), and flexibility (t2-Flex), while the proteases were described by four binary descriptors. The substrate descriptors were confined in the 'S block' and the protease descriptors in the 'P block' and were mean-centered and block-scaled to make them comparable. Separate proteochemometric models were constructed for rate of substrate cleavage (k_{cat}) and substrate binding (K_m) (see Section 4 for details). We first created linear models using only S and P blocks. These descriptors were later complemented by various square- and cross-terms (S^2, S × P, and S × S blocks) to identify non-linearities in substrate–protease interactions. The predictive ability of each model was evaluated by 7-fold

Table 1
Kinetic constants obtained with DEN-1–4 proteases using novel peptide substrates modified within the P1′–P4′ sequence with the general structure Abz–RRRR↓XXXX–nY–NH_2

Nr	P1′–P4′ sequence	DEN-1		DEN-2		DEN-3		DEN-4	
		k_{cat} (min^{-1})	K_m (μM)	k_{cat} (min^{-1})	K_m (μM)	k_{cat} (min^{-1})	K_m (μM)	k_{cat} (min^{-1})	K_m (μM)
1	APCN	2.38	11.43	0.24	5.41	0.67	3.87	2.86	4.94
2	HHGN	0.24	9.10	0.08	4.34	0.11	10.09	1.15	7.06
3	DDGN	0.27	170.98	0.07	79.59	NC[a]		0.54	89.51
4	SNSN	3.93	15.41	1.68	6.25	2.22	16.34	11.26	14.37
5	NWTN	0.87	13.73	0.30	6.02	0.39	9.37	2.27	10.41
6	GTVN	1.16	10.50	0.33	6.50	0.58	12.57	4.41	11.76
7	AGPN	1.54	12.56	0.41	4.18	0.62	10.83	4.47	19.15
8	SHCD	4.11	26.33	0.74	11.40	1.71	11.31	4.70	9.77
9	APGD	NC		0.04	19.55	NC		0.62	89.05
10	HWSD	0.15	17.19	0.07	14.86	NC		0.65	9.76
11	GGTD	4.29	49.72	0.43	12.93	1.48	18.64	3.51	12.21
12	AAVD	2.11	31.42	0.75	18.26	0.92	21.17	6.81	36.89
13	NDVD	0.36	106.17	0.10	43.17	NC		1.52	85.93
14	NLPD	0.59	36.84	0.22	16.71	0.32	53.49	2.02	33.48
15	HNCC	SI[b]		0.14	8.81	0.17	10.04	0.94	3.28
16	NACC	0.83	8.81	0.28	3.21	0.61	13.01	2.23	3.03
17	SWGC	SI		0.39	2.87	SI		2.39	4.79
18	AHSC	SI		0.26	5.63	SI		1.83	4.21
19	DSTC	0.46	15.23	0.07	12.81	0.20	11.12	0.84	23.08
20	SGVC	SI		0.98	4.26	1.75	19.72	4.54	7.06
21	GDPC	1.21	31.95	0.18	13.24	0.42	41.51	1.26	11.52
22	DGCH	0.77	22.89	0.09	7.02	0.17	20.84	0.47	9.05
23	ANGH	1.86	22.33	0.09	7.02	0.61	31.40	3.57	10.81
24	NPSH	0.37	13.41	0.08	5.18	0.11	26.92	1.35	21.47
25	GATH	4.95	17.43	0.94	4.66	1.42	18.86	4.69	7.79
26	HLVH	0.37	8.83	0.14	14.10	NC		1.08	16.29
27	HSPH	0.38	16.66	0.07	5.77	0.05	19.56	0.25	7.25
28	STPH	1.88	18.06	0.94	17.23	0.87	42.25	4.68	14.75
29	GWCF	NC		NC		NC		NC	
30	GLGF	3.19	12.26	0.90	4.33	1.41	16.60	8.11	21.17
31	DTSF	0.21	36.69	0.06	15.02	0.04	11.28	0.36	32.42
32	HPTF	SI		0.08	12.29	SI		1.34	24.69
33	ADTF	0.73	36.54	0.21	26.75	0.37	48.56	1.21	16.76
34	NSVF	1.84	24.75	0.29	10.84	0.46	28.86	NA[c]	
35	SAPF	8.17	20.03	2.00	5.33	2.87	22.21	11.48	13.28
36	ATCL	SI		0.22	7.31	SI		1.45	7.16
37	NGGL	4.13	21.04	0.10	7.17	1.30	25.42	4.58	9.57
38	HDSL	0.53	112.33	0.22	18.43	0.08	25.06	0.82	24.81
39	GSSL	2.21	15.38	0.48	10.44	1.07	25.92	2.69	6.92
40	SLTL	6.33	17.39	1.78	4.96	2.45	29.50	14.39	14.88
41	DHTL	0.34	54.56	0.07	28.85	NC		0.99	75.69
42	DNPL	0.29	33.75	0.07	17.31	0.17	75.26	1.07	87.96
43	SSC–[d]	SI		SI		2.97	7.43	5.96	8.48
44	HTG–	0.18	11.30	0.08	6.55	0.10	15.39	1.03	11.25
45	DLS–	0.54	53.82	0.14	19.65	0.14	69.33	0.77	21.78
46	NHT–	1.37	24.74	0.17	6.07	0.57	23.82	2.36	9.38
47	AWV–	1.69	11.49	0.27	2.24	0.86	18.19	5.09	8.51
48	GPP–	1.20	16.92	0.17	5.15	0.44	18.70	1.32	8.15
49	SAGM	9.98	25.5	1.86	3.44	3.34	24.86	16.93	22.65
50	SWPL	1.94	12.58	0.63	5.54	0.89	11.97	4.01	7.18
51	AGVL	3.60	23.85	0.34	2.19	0.63	10.59	3.32	10.84
52	GTGN	2.77	22.11	0.29	3.05	0.77	20.87	5.13	10.68
53	SAG–	16.91	28.39	2.54	3.61	4.11	20.90	27.67	15.20
54	AAGM	4.91	17.90	0.55	2.46	0.72	12.13	5.11	8.22
55	SAAM	9.18	16.34	1.56	3.33	1.80	13.85	13.15	14.33
56	SAGA	13.60	23.27	1.62	3.06	NA		13.34	10.49

[a] NC, minimal or no cleavage.
[b] SI, substrate inhibition.
[c] NA, not measured.
[d] –, denotes that there is no amino acid in the P4′ position.

cross-validation, here referred to as q^2, and a modified cross-validation with seven randomly formed groups of substrates, here referred to as q^2_{substr}, as well as by external predictions using the eight test set substrates, resulting in the q^2_{ext} estimate (see Section 4 for details). Modeling results are summarized in Table 2 for k_{cat} and in Table 3 for K_m. Linear models could approximate k_{cat} and K_m with good accuracy, explaining 83% of the variation in k_{cat} and 69% in K_m. Models also showed high predictive ability as assessed by the q^2 measure. However, q^2_{substr} and, especially, q^2_{ext} were substantially lower than the q^2, indicating that the linear models were

Table 2
Results of proteochemometric modeling of log k_{cat} data using different combinations of substrate and protease descriptors

Descriptor blocks	r^2	q^2	q^2_{substr}	q^2_{ext}
S, P	0.83	0.76	0.62	0.44
S, P, S^2	0.91	0.88	0.80	0.81
S, P, S^2, S × P	0.94	0.84	0.80	0.78
S, P, S^2, S × S	0.93	0.86	0.74	0.76

Table 3
Results of proteochemometric modeling of log K_m data using different combinations of substrate and protease descriptors

Descriptor blocks	r^2	q^2	q^2_{substr}	q^2_{ext}
S, P	0.69	0.59	0.54	0.40
S, P, S^2	0.75	0.68	0.65	0.54
S, P, S^2, S × P	0.80	0.68	0.66	0.57
S, P, S^2, S × S	0.78	0.66	0.63	0.45

poor predictors of activities from chemical properties of completely new substrates.

Squared-terms of substrate descriptors (S^2 block) were then added, which gave substantial improvements of the models. For the K_m model all three assessments of the predictive ability increased by 0.09–0.14 (Table 3), while the k_{cat} model improved even more (q^2 = 0.88, q^2_{substr} = 0.80, and q^2_{ext} = 0.81) (Table 2). Thus, if the linear model suffered in the ability to predict activities for new substrates, the predictive ability became excellent according to all assessments used herein by the inclusion of square-terms, S^2. One may therefore assume that strong non-linearities are present in the substrate property–activity relationships. For example, an optimal degree of hydrophilicity, size, or other property of substrate residues might exist that yields the most efficient molecular interactions with the proteases (*vide infra* for further interpretations of the models).

Addition of substrate–protease cross-terms (S × P block) improved the K_m model slightly, but had a negative influence on the predictive ability of the k_{cat} model. Moreover, including cross-terms for substrate descriptors (S × S block) also resulted in some decrease in the predictive ability for both K_m and k_{cat}. This lack of relevance of substrate cross-terms suggested that substrate prime side residues interact mostly in a non-cooperative manner, and, in turn, implied that the most preferable amino acids would be found separately for each prime side position of the substrates of the dengue proteases. Moreover, the lack of improvement of the k_{cat} model with the addition of the substrate–protease cross-terms showed that substrate cleavages by all the four dengue proteases are governed by closely similar processes, although the slight improvement of the K_m model indicated that the different dengue protease forms demonstrated some binding preferences among the different substrates.

The performances of k_{cat} and K_m models using S, P, and S^2 blocks are illustrated graphically in Figure 1. For only eight interaction pairs (i.e., 4% of the data) the predictions by the k_{cat} model were more than 0.5 logarithmic units wrong, as assessed by cross-validation and external validation (Fig. 1A). For the K_m model only four such mispredictions (2%) occurred (Fig. 1B).

2.4. Interpretation of models

With several highly predictive models for substrate binding and rate of substrate cleavage available, we next assessed the reliability of each model for interpretation by examining the differences between the goodness of fit and the predictive ability. Reliable interpretations can be attained if r^2 does not exceed the q^2 by more than 0.2–0.3 units; a larger difference is a sign of chance correlations to irrelevant descriptors, or of the presence of outliers in the data set.[37] As seen from Tables 2 and 3, the smallest margins between r^2 and q^2 are shown by the models comprising substrate descriptors, protease descriptors, and square-terms of substrate descriptors (i.e., S, P, and S^2 descriptor blocks); these models were accordingly used for interpretation of the roles of substrate and protease properties for substrate cleavage and binding. We also used the K_m model that in addition included the S × P descriptor block to investigate substrate selectivity between proteases of the four serotypes of dengue. For all three models, the difference $r^2 - q^2$ was in the range 0.03–0.12 while the difference $r^2 - q^2_{substr}$ varied over the narrow range of 0.10–0.14, which indicated that the model would be highly reliable for interpretations.

Interpretations of models were based on the analysis of partial least-squares projections to latent structures (PLS) regression equations. For a model using S, P, and S^2 descriptor blocks, the resulting regression equation can be expressed as follows:

$$y = \bar{y} + \sum_{P=1}^{4}(\mathrm{coeff}_P \times (x_P - \bar{x}_P) + \sum_{S=1}^{24}(\mathrm{coeff}_S \times (x_S - \overline{x_S}) + \mathrm{coeff}_{S^2} \times (x_S^2 - \overline{x_S^2}))$$

The sign and value of a PLS coefficient for a descriptor directly showed the influence of the represented property on the modeled activity, *y*. Interpretation of coefficients for square-terms is somewhat more intricate. Understanding of them is grounded on the fact that they become negative (after centering) when the values of original descriptors are close to their mean. On the other hand, square-terms get highly positive values if the values of the original descriptors are far from the mean. By comparing the coefficients for the square-term together with the original descriptor one may judge the degree of a non-linearity between the represented molecular property and the modeled activity.

Regression coefficients of the PLS models are represented graphically in Figure 2, where Figure 2A shows coefficients from the k_{cat} model and Figure 2B from the K_m model (Since a higher affinity for the substrate by the protease is indicated by a lower K_m constant, a positive coefficient for a descriptor in the K_m model denotes that the described molecular property correlates negatively with substrate binding, and vice versa).

As is revealed from the coefficients for the protease subtype descriptors of the k_{cat} model (Fig. 2A), the DEN-4 protease shows, in general, higher substrate cleavage rates, while the DEN-2 protease has the slowest rates. The negative coefficient for DEN-2 in the K_m model (Fig. 2B) indicates, on the other hand, that this protease subtype possesses overall higher affinity for the substrates. DEN-1 shows, on average, higher cleavage rates, while its substrate binding affinities are relatively low. Taken together, the data suggest that there is no apparent correlation between the coefficients for the protease descriptors for K_m and k_{cat}, a result that corroborates the lack of correlation between K_m and k_{cat} values mentioned above.

Comparison of the overall patterns for substrate descriptors in Figure 2A and B reveals major differences in the contribution of substrate properties to the protease k_{cat} and K_m activities. The K_m model has assigned substantially large PLS coefficients to several descriptors for each of the four prime site residue (Fig. 2B). This means that all prime side residues influenced the binding of substrates by the dengue proteases. By contrast, for the k_{cat} model, some descriptors for the P1′ and P2′ positions gained considerably large PLS coefficients, while at the P3′ and P4′ positions, the coefficients for all descriptors were minor (Fig. 2A). Accordingly, only the P1′ and P2′ positions substantially influence the rate of substrate cleavage.

Further analysis of the PLS coefficients at the P1′ position of the k_{cat} model revealed large negative values for most of the substrate square-terms. This result implies that the optimal amino acid(s) for the P1′ position reside within the modeled amino acids' 'physicochemical property space', and that large deviation from the center of this space negatively influence substrate turnover. For the original descriptors (i.e., non-squared descriptors), large negative coefficients are assigned to size and rigidity. Thus, these two properties mainly cause a reduction in the rate of substrate cleavage. By contrast, the coefficient for the polarity descriptor (zz_3) is close to zero,

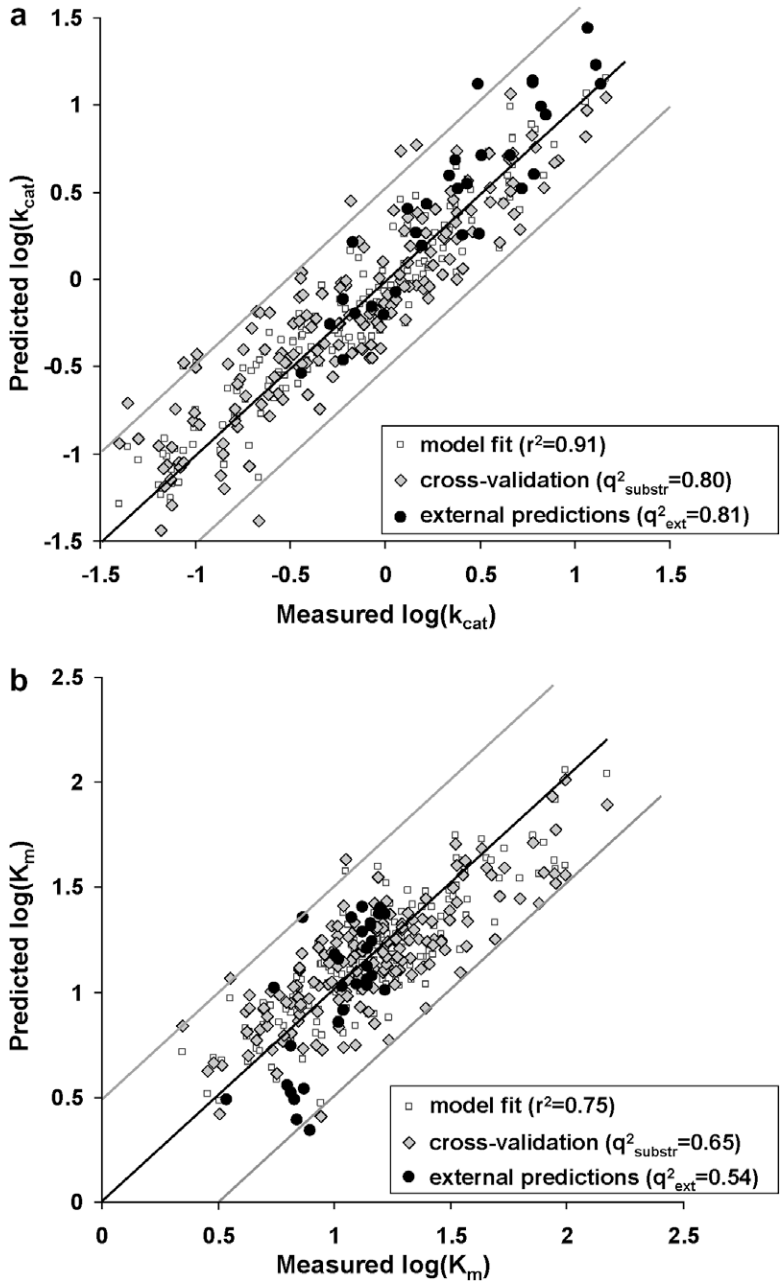

Figure 1. Correlation of predicted versus observed log(k_{cat}) (A) and log(K_m) (B) values derived from the proteochemometric models using substrate and protease descriptors, and square-terms of substrate descriptors. Predictive ability is assessed by cross-validation leaving out 1/7 of substrates at a time (gray diamonds) and by external validation (black circles). Shown is also model fit (small unfilled squares). Indicated by the oblique gray lines is the plot area where prediction errors do not exceed 0.5 logarithmic units.

indicating that broad variations are acceptable at P1′ for high k_{cat} values. The hydrophilicity and flexibility descriptors (i.e., zz_1 and t2-Flex) yielded some negative correlation with k_{cat}, but the effects are nonlinear as is indicated by the large coefficients for their square-terms. Accordingly, very hydrophobic amino acids or amino acids that lack any flexible features are not advantageous at P1′. Detailed analysis of the various molecular properties at the P1′ position identified Ser as the best amino acid for high k_{cat}. Serine ensured, on average, 3-fold and 5-fold higher cleavage rates than Ala and Gly, respectively, the two amino acids closest to Ser for achieving high k_{cat} (These conclusions are in fair agreement with earlier findings[27]).

The pattern of coefficients for descriptors at the P2′ position (k_{cat} model) was completely different. Here, the largest positive coefficients were given to square-terms of the hydrophilicity and flexibility (i.e., zz_1 and t2-Flex) descriptors, whereas the most negative influence arose from rigidity features (i.e., t1-Rig) of the residue. By comparing the values of all descriptors and square-terms for all ten amino acids present in the P2′ among our substrates, we found that flexible or small amino acids (e.g., Ala, Gly, Leu) should be beneficial at this position. Moreover, as indicated by the positive coefficient for the charge descriptor, the acidic Asp is inferior to Asn at this position. Among the identified less favorable residues at P2′ were Thr and Trp. This result was unexpected since these amino acids are present at the P2′ position in some native cleavage sites of the dengue viruses. However, our test substrates containing Thr and Trp at the P2′ position indeed showed relatively low k_{cat} activities, as predicted by the model, and thus supported these interpretations.

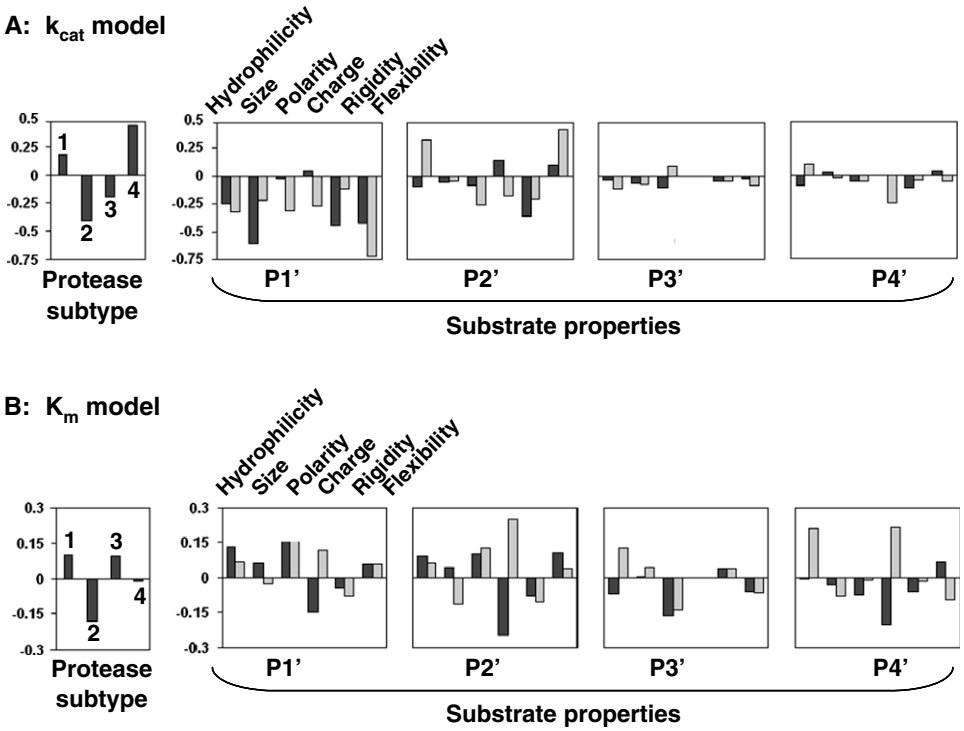

Figure 2. Regression coefficients of proteochemometric models of log(k_{cat}) (A) and log(K_m) (B) data. In each panel, shown are regression coefficients for the four protease subtype descriptors (on the left) and descriptors of six properties for the four substrate residues, P1′–P4′. Solid black bars represent regression coefficients for original descriptors (Note that negative values of hydrophilicity descriptors are seen for hydrophobic amino acids whereas positive values are seen for hydrophilic amino acids; negative values of size descriptor characterize small, compact amino acids, while positive values are seen among large, bulky amino acids; for polarity descriptor, a negative value indicates that the amino acid is electrophilic, whereas a positive value indicates that it is electronegative. The charge descriptor reflects the charge of the amino acid. Positive values of the rigidity and flexibility descriptors indicate the presence of, respectively, a high fraction of rigid and flexible fragments in an amino acid, while negative values indicate the lack of such fragments.) and gray bars represent regression coefficients for their square-terms.

The coefficients for descriptors at the P3′ and P4′ positions were small, thus showing that the model did not identify any appreciable correlations between properties of these positions and the k_{cat} activity.

However, the patterns were very different for the coefficients of the K_m model. Here all four P′ positions gained considerably large PLS coefficients for one or other of the descriptors (Fig. 2B). The highest absolute values of coefficients were obtained by the charge descriptors of the P2′ and P4′ sites. The negative values of these coefficients identified the acidic Asp and Glu residues as unfavorable for high protease binding affinity. On the other hand, the square-terms of these two descriptors attained positive coefficients, which showed that correlation between charge and K_m is not linear over the whole range of values (e.g., His is not much preferred over a neutral amino acid at P4′). A large positive coefficient is also assigned to the square-term of the hydrophilicity descriptor at P4′, indicating that extremely hydrophobic or hydrophilic amino acids have here a negative impact on substrate binding. Although the coefficients for polarity and flexibility are rather small at P4′, these data suggest that Ala, Gly, and Cys (and perhaps Pro) could improve the K_m activity.

For the P3′ position, the most important were descriptors of polarity and hydrophilicity, and also their square-terms. Careful analysis of these coefficients and the descriptor values for all amino acids in the work set indicated some preference for Cys over the other amino acids including Gly, Ser, Thr, Val, and Pro, the five amino acids that are present at P3′ in native substrate cleavage sites of dengue. Thus, the model finds that P3′ in native substrates is not optimal for protease binding affinity.

At the P2′ position, the charge descriptor gave an important contribution to K_m, explaining the low activity of substrates containing aspartic acid at this position. Several other P2′ descriptors (zz_1, zz_3, and t2-Flex) and their cross-terms obtained positive coefficients. These combined results showed that hydrophilicity, electronegativity, and flexibility tended to correlate with K_m in a nonlinear manner. By comparing the predicted K_m values when all ten amino acids were tried at P2′, we can conclude that this position allows a broad range of amino acids, with the most favorable being Trp, closely followed by His, Thr, Ala, and Gly.

The patterns of coefficients for the P1′ and P2′ positions are quite similar (Fig. 2B). However, when interpreting the requirements at these positions one should take into account that we had a wider selection of amino acids for the P2′ position than for the P1′ position; for the latter, the chemical space covered was more compact (Preliminary studies suggested that hydrophobic and electrophilic amino acids were not appropriate at P1′ for obtaining cleavable substrates). Despite these limitations, our interpretation for the P1′ position revealed that Ala, Gly, and Ser, which were beneficial for high k_{cat} activity, also provided good binding affinity. On the other hand, the coefficients for charge and polarity descriptors revealed that the acidic and electronegative amino acid Asp was the least appropriate at the P1′ position for tight enzyme binding.

To complete the interpretations, we investigated the PLS coefficients for substrate–protease cross-terms in the K_m model built on the S, P, S^2, and S × P descriptor blocks, which was also the model that showed the highest q^2_{substr} and q^2_{ext} values. Three of the four highest PLS coefficient values were here given to the cross-terms formed between the DEN-2 descriptor and the flexibility descriptors at P3′ and P2′, and the rigidity descriptor at P1′. Hence, we may conclude that the DEN-2 protease deviates most from other subtypes in its substrate binding preferences. For the P3′ position,

particularly large selectivity differences would arise, according to these results, for Cys-containing substrates, and such a substrate would be less favorable for the DEN-2 protease. A high PLS coefficient was also given to the cross-term of the DEN-3 descriptor with the charge descriptor at P4′, indicating that Asp has a less negative influence on the binding of the substrate to DEN-3 than for the other protease subtypes.

Because the majority of cross-terms obtained very low PLS coefficients, they played minor roles in the modeling. The negligible increase of r^2 and q^2 values of PLS model upon including of cross-terms supports this view. Thus, we may conclude that DEN-1–4 proteases share similar substrate recognition mechanisms. This implies that a protease inhibitor may be designed which is simultaneously capable of targeting all four subtypes of the dengue virus.

3. Conclusions

Rather than applying a large combinatorial library approach, we elected here to analyze data of individual substrates, evaluating 6–10 of the most promising amino acids for each of the P1′–P4′ residues. To make the task affordable, we applied statistical molecular design. By using D-optimal design we created a balanced and representative substrate library of a reasonable size. For the modeling, we encoded the varying P1′–P4′ positions of substrates by descriptors of amino acids, representing essentially all physico-chemical properties that could be important for substrate–protease interactions. The proteochemometric models included concomitantly all substrates and all four dengue proteases. Models where thoroughly validated for their predictive ability (q^2_{substr} for the k_{cat} model being 0.80 and for the K_{m} model, 0.65) and reliability of interpretations (the difference $r^2 - q^2_{\text{substr}}$ being as low as 0.11 and 0.10 for, respectively, the k_{cat} and K_{m} models). We may thus conclude that for all statistical aspects of model validation, these models are highly reliable.

It is common to characterize enzyme kinetics using $k_{\text{cat}}/K_{\text{m}}$ ratios only. However, if one aims to find a peptide with both a low K_{m} and low k_{cat}, which might serve as a lead peptide in the design of a protease inhibitor, analysis of the $k_{\text{cat}}/K_{\text{m}}$ ratio is not sufficient. From the raw data on the kinetic activities of the peptide substrate series synthesized herein, it was evident that there was no relationship between substrate affinity and cleavage rate. The proteochemometric modeling revealed that the values of k_{cat} and K_{m} of the dengue proteases are affected by different properties of the substrate amino acids. Thus, the proteases demonstrated a high rate of cleavage for peptides having small amino acids (Ser, Gly, Ala) at the P1′ position and small or flexible amino acids at the P2′ position, while the P3′ and P4′ positions had little effect on substrate turnover. By contrast, all four P′ positions of the substrate significantly influenced K_{m} of the proteases. Thus, for high affinity (low K_{m}), a P1′ amino acid should be small and moderately hydrophilic, while acidic residue is highly unfavorable. For the P2′ position, the most favorable was Trp; however, this position allowed a broader diversity of amino acids. For the P3′ position, Cys was more beneficial than the amino acids present in native cleavage sites of the dengue polyproteins. For the P4′ position, the best affinity would be obtained with Ala, Gly, or Cys.

These interpretations support largely the results of our previous study, which provided measured k_{cat} and K_{m} data for combinations of prime and non-prime side sequences of natural cleavage sites of DEN-2.[27] However, our results are not directly comparable to the recent study by Li et al. who investigated pools of substrates of DEN-1–4 proteases.[26] In the latter study, the P′ site specificities of dengue substrates were determined by comparing initial velocity values for a single concentration of peptide mixtures, and Ser was found to be the optimal amino acid for P1′ and P3′. Our results confirm that Ser at P1′ is highly beneficial for both kinetic constants, but this was not the case at P3′. Our substrates that included Ser at P3′ did not show superior k_{cat} or K_{m} activity. However, one should keep in mind that each measurement in the study by Li et al. was obtained using a mixture of $19 \times 19 \times 19$ amino acid combinations at three of the four P′ side positions. Artifacts could then arise because for some peptides, a bond hydrolysis at a position other than the predicted scissile bond could occur (i.e., at the bond after the two arginines at P2–P1).

To our knowledge, our study is the first to report a quantitative analysis for the contributions of physico-chemical properties of residues of peptide substrates with respect to the kinetics of dengue proteases. Our models might be used to modify the substrates by natural or synthetic amino acids to optimize separately k_{cat} and K_{m} constants for the DEN-1-4 proteolytic enzymes. Design and evaluation of high affinity uncleavable peptides which might serve as templates for peptidomimetic inhibitors of dengue should be a task of further studies.

4. Experimental section

4.1. Synthesis of peptides

Internally quenched fluorescent substrates were based on the Abz (*o*-aminobenzoic acid)/3-nitrotyrosine (Tyr(3-NO_2) or nY) pair and were prepared by solid-phase synthesis using an automated multiple peptide synthesizer (MultiPep; Intavis Bioanalytical Instruments AG, Koeln, Germany) using the automated standard protocol optimized for Fmoc chemistry. Reagents were purchased from Fluka (Sigma–Aldrich, St. Louis, MO, U.S.A.), Applied Biosystems (Foster City, CA, U.S.A.), Bachem (Bubendorf, Switzerland) and Novabiochem (Calbiochem-Novabiochem AG, Laufelfingen, Switzerland). The following amino acid derivatives were used: Fmoc-Ala-OH (where Fmoc is fluoren-9-ylmethoxycarbonyl), Fmoc-Arg(Pbf)-OH (where Pbf is 2,2,4,6,7-pentamethyldihydrobenzofuran-5-sulphonyl), Fmoc-Asn(Trt)-OH (where Trt is trityl), Fmoc-Asp(O*t*-Bu)-OH (where *t*-Bu is *t*-butyl), Fmoc-Cys(Trt)-OH, Fmoc-His(Trt)-OH, Fmoc-Gln(Trt)-OH, Fmoc-Gly-OH, Fmoc-Leu-OH, Fmoc-Lys(Boc)-OH (where Boc is *t*-butoxycarbonyl), Fmoc-Met-OH, Fmoc-Phe-OH, Fmoc-Pro-OH, Fmoc-Ser(*t*-Bu)-OH, Fmoc-Thr(*t*-Bu)-OH, Fmoc-Trp(Boc)-OH, Fmoc-Val-OH, Fmoc-Tyr(3-NO_2)-OH, and Boc-Abz-OH. PyBOP (benzotriazole-1-yl-oxy-tris-pyrrolidino-phosphonium hexafluorophosphate) was used as the activating reagent and Rink Amide MBHA resin (capacity 0.64 mmol/g) as the polymeric support [Rink Amide MBHA resin is 4-(2′,4′-dimethoxyphenyl-Fmoc-aminomethyl)-phenoxyacetamido-norleucyl-4-methylbenz-hydrylamine resin]. Peptides were characterized by HPLC and their structures were confirmed by MS. Analytical HPLC was performed on a Waters (Milford, MA, U.S.A.) system (Millenium32 workstation, 2690 Separation Module, 996 Photodiode Array Detector) equipped with a Vydac RP C18 90 Å reversed-phase column (2.1 mm × 250 mm). The purity of raw peptides according to HPLC was above 80%, and they were used for kinetic experiments after freeze-drying. Molecular mass measurements were performed on a PerkinElmer (PerkinElmer Life and Analytical Sciences, Boston, MA, U.S.A.) instrument PE SCIEX API 150EX with TurboIonSpray ion source. Freeze-drying was carried out at 0.01 bar (1 bar = 100 kPa) on a Lyovac GT2 freeze-dryer (Steris Finn-Aqua, Tuusula, Finland) equipped with a Trivac D4B (Leybold Vacuum GmbH, Cologne, Germany) vacuum pump and a liquid nitrogen trap. All other chemicals were reagent grade from Sigma.

4.2. Expression and purification of dengue proteases

pTrcHis plasmids containing a recombinant NS2B(H)-NS3pro sequence from one of the dengue virus proteases 1–4 were transformed into *Escherichia coli* C41(DE3) as described previously.[27] Transformants were grown in Luria broth (LB) medium supplemented with ampicillin (100 μg/mL) at 37 °C. When OD_{600} reached about 0.5, isopropyl-1-thio-β-D-galactopyranoside was added to a final concentration of 0.1 mM, and the culture was grown at 37 °C for 8 h. Cells were harvested by centrifugation (5000 *g*, 10 min, 4 °C), resuspended in 20 mL of lysis buffer A (100 mM Tris–HCl, pH 7.5, 300 mM NaCl), and lysed with a sonicator (MPS) 6 V, 5 × 30 s. The lysate was sedimented by centrifugation (10,000*g*, 30 min, 4 °C), and the pellet with inclusion bodies was washed two times with lysis buffer containing 1% Triton X-100.

The above pellet was then suspended in 15 mL of denaturing buffer B (100 mM Tris–HCl, pH 8.0, 300 mM NaCl, 8 M urea), homogenized with a Ultra-Turrax T25 tissue disperser (Labassco) and centrifuged (10,000*g*, 30 min, 4 °C), whereafter the supernatant was loaded on a Hitrap chelating column (Pharmacia) equilibrated with denaturing buffer. The column was washed with 10 column volumes of denaturing buffer containing 20 mM imidazole and eluted at a flow rate of 0.5 mL/min with denaturing buffer containing 200 mM imidazole. Fractions of 1 mL were collected, and aliquots were analyzed for the presence of NS2B(H)-NS3pro by sodium dodecyl sulfate (SDS)–polyacrylamide gel electrophoresis (PAGE) on 15% polyacrylamide gels.

Peak fractions were pooled and loaded on a Superdex 200 HR 10/30 gel filtration column (Pharmacia). The column was eluted with denaturing buffer at a flow rate of 0.3 mL/min and the fractions containing NS2B(H)-NS3pro, as analyzed by SDS–PAGE, were pooled and diluted with the same buffer to 0.5 mg/mL. Refolding of the protein was initiated by stepwise dialysis of 1 mL samples with a dialysis tubing (cutoff, 8 kDa) at 4 °C against three changes of 100 mM Tris–HCl, pH 8.0–300 mM NaCl (200 mL), and one change against 200 mL of 100 mM Tris–HCl, pH 9.0–50 mM NaCl (buffer C). The dialysate was centrifuged (10,000*g*, 10 min, 4 °C) and the protein concentration was determined with a Bradford protein assay kit (Bio-Rad). Preparations of the NS2B(H)-NS3pro protein were stored at −20 °C in 100 mM Tris–HCl, pH 9.0–50 mM NaCl-50% glycerol.

4.3. Kinetic characterization of peptides on dengue proteases

Fluorogenic assays were carried out with the 56 substrates prepared above by using a POLARstar OPTIMA 96 well plate reader (BMG Labtech GmbH). Substrates cleavage over time was followed by monitoring the emission at 420 nm upon excitation at 320 nm. Final reaction volumes were 100 μL and contained 50 mM Tris–HCl, pH 9.0 and 20% glycerol and the incubation temperature was 37 °C. Enzyme concentrations were varied over the range of 150–200 nM, depending on the substrate used. The substrate concentrations were varied between 1 and 100 μM (12 dilution points for each substrate). For each substrate concentration the initial rates of enzyme cleavage were computed using "first order rate equation with offset" equation as implemented in program GraFit, version 5 (Sigma). From the obtained initial rates, the K_m and k_{cat} constants were computed according to Michaelis–Menten enzyme kinetics equation using the GraFit version 5 program. Measurements were repeated at least three times for each protease–substrate pair.

4.4. Numerical description of proteases and substrates

For the sake of the proteochemometric modeling, the measured K_m and k_{cat} activity data were correlated to quantitative descriptions of both the peptide substrates and the proteases. Proteases were described by four indicator variables, each representing the protease from one serotype, DEN-1–4. For example, if the activity measurement was done using the DEN-1 protease, the values of the four descriptors were 1, 0, 0, and 0; if the activity measurement was done using the DEN-2 protease, the values of the four descriptors were 0, 1, 0, and 0, and so on.

The substrates were characterized by descriptors for each varied amino acid within the library, representing the following six molecular properties: hydrophilicity (represented by zz_1 descriptor), size (zz_2), polarity (zz_3), charge (C7.4), rigidity (t1-Rig), and flexibility (t2-Flex). The zz-scales, zz_1-zz_3 were proposed earlier by Sandberg et al.[41] and had been obtained by principal component analysis (PCA)[42] of 26 measured and computed physicochemical properties of 87 natural and artificial amino acids. Zz_1-zz_3 are the three first principal components; accordingly, they represent the largest variations within all analyzed physico-chemical properties and are orthogonal (i.e., uncorrelated). In the zz_1-scale, hydrophobic amino acids are represented by negative values and hydrophilic amino acids by positive values. In zz_2, negative values represent small, compact amino acids while positive values represent large, bulky amino acids. In zz_3, a negative value indicates that the amino acid is electrophilic whereas a positive value indicates that it is electronegative. The C7.4 descriptor was proposed by Gottfries[43] and represents the proportion of the amino acid side chains that are ionized at pH = 7.4 (i.e., for anionic Asp and Glu, C7.4 = −1, for cationic His, +0.5 and for Arg and Lys, +1, and for all other amino acids C7.4 = 0). Descriptors t1-Rig and t2-Flex were also calculated by Gottfries[43] by applying PCA to a set of constitutional descriptors of amino acids, which included number of rings, rigid bonds and rotatable bonds, rigid fragments, flexible and partially flexible chains, length of the longest flexible chain, and so forth; t1-Rig and t2-Flex are first two principal components of these and are accordingly uncorrelated to each other. In this way, each of the P1′–P4′ positions was described by six descriptors of amino acid properties, yielding totally 24 substrate descriptors (protease descriptors will be referred to as P block, and substrate descriptors as S block). During modeling, we found that the relationships between activities and substrate descriptors were only partially linear. To account for this discrepancy, we calculated squares of mean-centered substrate descriptors, and used these 24 square-terms as an additional descriptor block (here termed S^2 block). Moreover, we also investigated possible cooperation of substrate and protease properties. For this purpose substrate–protease and substrate–substrate cross-terms were computed by multiplication of mean-centered descriptors. Thus, two additional blocks were obtained: S × P comprising 24 × 4 = 96 descriptors and S × S comprising 24 × 23/2 = 276 descriptors.

4.5. Data pre-processing

All descriptors were mean-centered and scaled to unit variance prior to further use. Block-scaling was also applied to prevent the situation that the large amount of cross-terms would dominate the correlations and the few protease descriptors then would become almost indiscernible. In block scaling, all variables of each block are multiplied by some constant value, the block scaling factor. Herein, for each of the five descriptor blocks (i.e., P, S, S^2, S × P, and S × S) the scaling factor was assigned as $1/\sqrt{k}$, where k is the number of descriptors in the given block. In this way, we ensured that all blocks were equally weighted in the modeling. The two activities, K_m and k_{cat}, were logarithmically transformed and mean-centered prior to applying further calculations.

4.6. Correlation by partial least-squares projections to latent structures

Descriptors were correlated to the logarithms of the K_m and k_{cat} activities of the 48 work set peptides by partial least-squares projections to latent structures (PLS)[44] using the Unscrambler 9.7 software (CAMO Software AS, Norway). PLS is a widely used method for finding a quantitative relationship between a set of descriptors (X data) and one or several responses (Y data). This is achieved by simultaneously projecting the **X** and **Y** matrices onto lower dimensionality variable space (PLS components) with an additional constraint to maximize the covariance between projections of **X** and **Y**. For each response, PLS derives a regression equation, where regression coefficients show the direction and magnitude of the influence of descriptors on the response (for detailed descriptions see Refs. 45 and 46).

4.7. Validation of models

We used thorough validations to find the optimal complexity (i.e., number of components) of the PLS models and to assess their reliability for interpretations and predictions. The goodness of fit of a PLS model was assessed by r^2, the fraction of the explained variance of the response. The predictive ability was characterized by q^2, the fraction of the predicted variance of the response assessed by cross-validation.[45]

In the current study, cross-validation was performed in two ways. First, the dataset was randomly divided into seven groups. In this way, we assessed the predictive ability for new substrate–protease combinations. Cross-validation groups were thereafter rearranged so that all four activity measurements for each of 48 substrates were assigned to the same cross-validation group. This was done to avoid overly optimistic assessments of the predictive ability in case the four protease subtypes shared similar substrate interaction profiles (Results from conventional cross-validation are here referred to as q^2, and results from modified variant of cross-validation as q^2_{substr}). Finally, we also estimated the external predictive ability of models by performing predictions for the eight test set substrates (The assay data for the test set substrates were obtained independently using the same assay method as for the work set substrate and are given in Table 1). The thus obtained q^2_{ext} estimate differs from q^2_{substr} essentially by the fact that P1′–P4′ of the test set peptides resemble native cleavage sites of DEN-1–4 polyproteins and, as a result, these peptides generally show higher activities than the average for the work set. Thus, q^2_{ext} measure is preferable for the assessment of the ability of models to generalize towards high activities.

Acknowledgments

We thank Ewelina Fogelström for technical assistance. Support was obtained by SIDA-Swedish Research Links (348-2004-5993) and the Swedish VR (04X-05957) and by basic science grant BRG 49800008 (to G.K.) and a Royal Golden Jubilee scholarship from the Thailand Research Fund (TRF).

References and notes

1. Halstead, S. B. *Science* **1988**, *239*, 476.
2. Jacobs, M. G.; Young, P. R. *Curr. Opin. Infect. Dis.* **1998**, *11*, 319.
3. Monath, T. P. *Proc. Natl. Acad. Sci. U.S.A.* **1994**, *91*, 2395.
4. Gubler, D. J.; Clark, G. G. *Emerg. Infect. Dis.* **1995**, *1*, 55.
5. Guzman, M. G.; Kouri, G. *Lancet Infect. Dis.* **2002**, *2*, 33.
6. Gubler, D. J. *Clin. Microbiol. Rev.* **1998**, *11*, 480.
7. Gubler, D. J. *Trends Microbiol.* **2002**, *10*, 100.
8. Edelman, R. *Clin. Infect. Dis.* **2007**, *45*(Suppl. 1), S56.
9. Chambers, T. J.; Hahn, C. S.; Galler, R.; Rice, C. M. *Annu. Rev. Microbiol.* **1990**, *44*, 649.
10. Stadler, K.; Allison, S. L.; Schalich, J.; Heinz, F. X. *J. Virol.* **1997**, *71*, 8475.
11. Falgout, B.; Pethel, M.; Zhang, Y. M.; Lai, C. J. *J. Virol.* **1991**, *65*, 2467.
12. Zhang, L.; Mohan, P. M.; Padmanabhan, R. *J. Virol.* **1992**, *66*, 7549.
13. Ramachandran, M.; Sasaguri, Y.; Nakano, R.; Padmanabhan, R. *Methods Enzymol.* **1996**, *275*, 168.
14. Preugschat, F.; Yao, C. W.; Strauss, J. H. *J. Virol.* **1990**, *64*, 4364.
15. Chambers, T. J.; Weir, R. C.; Grakoui, A.; McCourt, D. W.; Bazan, J. F.; Fletterick, R. J.; Rice, C. M. *Proc. Natl. Acad. Sci. U.S.A.* **1990**, *87*, 8898.
16. Preugschat, F.; Strauss, J. H. *Virology* **1991**, *185*, 689.
17. Arias, C. F.; Preugschat, F.; Strauss, J. H. *Virology* **1993**, *193*, 888.
18. Lin, C.; Amberg, S. M.; Chambers, T. J.; Rice, C. M. *J. Virol.* **1993**, *67*, 2327.
19. Lobigs, M. *Proc. Natl. Acad. Sci. U.S.A.* **1993**, *90*, 6218.
20. Teo, K. F.; Wright, P. J. *J. Gen. Virol.* **1997**, *78*, 337.
21. Li, H.; Clum, S.; You, S.; Ebner, K. E.; Padmanabhan, R. *J. Virol.* **1999**, *73*, 3016.
22. Clum, S.; Ebner, K. E.; Padmanabhan, R. *J. Biol. Chem.* **1997**, *272*, 30715.
23. Falgout, B.; Miller, R. H.; Lai, C. J. *J. Virol.* **1993**, *67*, 2034.
24. Chambers, T. J.; Grakoui, A.; Rice, C. M. *J. Virol.* **1991**, *65*, 6042.
25. Wengler, G.; Czaya, G.; Farber, P. M.; Hegemann, J. H. *J. Gen. Virol.* **1991**, *72*, 851.
26. Li, J.; Lim, S. P.; Beer, D.; Patel, V.; Wen, D.; Tumanut, C.; Tully, D. C.; Williams, J. A.; Jiricek, J.; Priestle, J. P.; Harris, J. L.; Vasudevan, S. G. *J. Biol. Chem.* **2005**, *280*, 28766.
27. Niyomrattanakit, P.; Yahorava, S.; Mutule, I.; Mutulis, F.; Petrovska, R.; Prusis, P.; Katzenmeier, G.; Wikberg, J. *Biochem. J.* **2006**, *397*, 203.
28. Yin, Z.; Patel, S. J.; Wang, W. L.; Wang, G.; Chan, W. L.; Rao, K. R.; Alam, J.; Jeyaraj, D. A.; Ngew, X.; Patel, V.; Beer, D.; Lim, S. P.; Vasudevan, S. G.; Keller, T. H. *Bioorg. Med. Chem. Lett.* **2006**, *16*, 36.
29. Chanprapaph, S.; Saparpakorn, P.; Sangma, C.; Niyomrattanakit, P.; Hannongbua, S.; Angsuthanasombat, C.; Katzenmeier, G. *Biochem. Biophys. Res. Commun.* **2005**, *330*, 1237.
30. Lapinsh, M.; Prusis, P.; Gutcaits, A.; Lundstedt, T.; Wikberg, J. E. S. *Biochim. Biophys. Acta* **2001**, *1525*, 180.
31. 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.
32. Lapinsh, M.; Prusis, P.; Uhlén, S.; Wikberg, J. E. S. *Bioinformatics* **2005**, *21*, 4289.
33. Lapinsh, M.; Veiksina, S.; Uhlén, S.; Petrovska, R.; Mutule, I.; Mutulis, F.; Yahorava, S.; Prusis, P.; Wikberg, J. E. S. *Mol. Pharmacol.* **2005**, *67*, 50.
34. Prusis, P.; Uhlen, S.; Petrovska, R.; Lapinsh, M.; Wikberg, J. E. S. *BMC Bioinformatics* **2006**, *7*, 167.
35. Mandrika, I.; Prusis, P.; Yahorava, S.; Shikhagie, M.; Wikberg, J. E. S. *Protein Eng. Des. Sel.* **2007**, *20*, 301.
36. Kontijevskis, A.; Prusis, P.; Petrovska, R.; Yahorava, S.; Mutulis, F.; Mutule, I.; Komorowski, J.; Wikberg, J. E. S. *PLoS Comput. Biol.* **2007**, *3*, e48.
37. Eriksson, L.; Johansson, E. *Chemom. Intell. Lab.* **1996**, *34*, 1.
38. Lundstedt, T. E.; Seifert, E.; Abramo, L.; Thelin, B.; Nystrom, A.; Pettersen, J.; Bergman, R. *Chemom. Intell. Lab.* **1998**, *42*, 3.
39. De Aguiar, P. F.; Bourguignon, B.; Khots, M. S.; Massart, D. L.; Phan-Than-Luu, R. *Chemom. Intell. Lab.* **1995**, *2*, 199.
40. Eriksson, L.; Johansson, E.; Kettaneh-Wold, N.; Wikström, N.; Wold, S. *Design of Experiments, Principles and Applications*; Umetrics AB: Umeå, 2000.
41. Sandberg, M.; Eriksson, L.; Jonsson, J.; Sjöström, M.; Wold, S. *J. Med. Chem.* **1998**, *41*, 2481.
42. Wold, S.; Esbensen, K.; Geladi, P. *Chemom. Intell. Lab.* **1987**, *2*, 37.
43. Gottfries, J. *Chemom. Intell. Lab.* **2006**, *83*, 148.
44. Wold, S.; Sjöström, M.; Eriksson, L. *Chemom. Intell. Lab.* **2001**, *58*, 109.
45. Wold, S. PLS for Multivariate Linear Modeling. In *Chemometric Methods in Molecular Design*; van de Waterbeend, H., Ed.; VCH: Weinheim, Germany, 1995; pp 195–218.
46. Geladi, P.; Kowalski, B. R. *Anal. Chim. Acta* **1986**, *185*, 1.