Computational Proteomics Analysis of HIV-1 Protease Interac-tome

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · JULY 2007

Impact Factor: 2.63 · DOI: 10.1002/prot.21415 · Source: PubMed

3 AUTHORS, INCLUDING:

Jarl ES Wikberg
Uppsala University
301 PUBLICATIONS **9,517** CITATIONS
SEE PROFILE

Jan Komorowski
Uppsala University
179 PUBLICATIONS **7,519** CITATIONS
SEE PROFILE

PROTEINS: Structure, Function, and Bioinformatics 68:305–312 (2007)

Computational Proteomics Analysis of HIV-1 Protease Interactome

Aleksejs Kontijevskis,[1,2] Jarl E. S. Wikberg,[2] and Jan Komorowski[1]*
[1]*The Linnaeus Centre for Bioinformatics, Uppsala University, S-75124 Uppsala, Sweden*
[2]*Department of Pharmaceutical Biosciences, Uppsala University, S-75124 Uppsala, Sweden*

ABSTRACT **HIV-1 protease is a small homodimeric enzyme that ensures maturation of HIV virions by cleaving the viral precursor Gag and Gag-Pol polyproteins into structural and functional elements. The cleavage sites in the viral polyproteins share neither sequence homology nor binding motif and the specificity of the HIV-1 protease is therefore only partially understood. Using an extensive data set collected from 16 years of HIV proteome research we have here created a general and predictive rule-based model for HIV-1 protease specificity based on rough sets. We demonstrate that HIV-1 protease specificity is much more complex than previously anticipated, which cannot be defined based solely on the amino acids at the substrate's scissile bond or by any other single substrate amino acid position only. Our results show that the combination of at least three particular amino acids is needed in the substrate for a cleavage event to occur. Only by combining and analyzing massive amounts of HIV proteome data it was possible to discover these novel and general patterns of physico-chemical substrate cleavage determinants. Our study is an example how computational biology methods can advance the understanding of the viral interactomes. Proteins 2007;68:305–312.** **© 2007 Wiley-Liss, Inc.**

Key words: viral proteomics; bioinformatics; protein–peptide interactions; HIV-1 protease specificity; viral complexity

INTRODUCTION

The high rate of replication of HIV and the development of drug resistance are probably the most challenging problems in modern AIDS therapy. It is estimated that during the latent or steady stage of infection over 10^{10} new cells become infected each day in a typical HIV-infected patient and that every possible single-point mutation along the viral genome arises 10^4–10^5 times daily.[1] This results in a large number of viable viruses harboring multiple mutations in their structural and functional proteins.[1,2] To replicate successfully, the virus requires its HIV-1 protease to cleave viral precursor Gag and Gag-Pol polyproteins into structural and functional elements, which ensures the maturation of new virions.[3,4] HIV-1 protease has served as an attractive target for design of therapeutic inhibitors to prevent formation of infectious HIV forms.[5] However, because of the polymorph nature of HIV-1 protease, protease inhibitor resistant strains of HIV emerge in any HIV infected person within certain amount of time. Therefore, a successful development of potent protease inhibitors capable to adapt to many HIV-1 protease variants requires a deep understanding of its protease-substrate interactions.

HIV-1 protease recognizes an octapeptide sequence represented by P_4-P_3-P_2-$P_1{\downarrow}P_1'$-P_2'-P_3'-P_4', where "↓" denotes a scissile bond, P_m denotes an amino acid at the substrate N-terminus and P_m'—an amino acid at the substrate C-terminus. The cleavage sites in the viral polyproteins do not share any obvious sequence homology or binding motif and therefore the specificity of the HIV-1 protease is only partially understood.[3,6] Traditionally, HIV-1 protease substrates have been roughly classified into three major groups according to the amino acids in the P_1 and P_1' positions: (i) aromatic ↓ Proline; (ii) hydrophobic ↓ hydrophobic; and (iii) other cleavage sites.[3] Such classifications are typically based on the comparative analysis of relatively small sets of substrates, such as those represented by natural cleavage site peptides in the Gag and Gag-Pol polyproteins, with mono or double amino acid substitutions.[3]

Various prediction methods have been applied with some success to determine substrate cleavability by HIV-1 protease.[7–17] Although these methods were able to predict cleavability, all of them were in fact based on the analysis of the data set that consisted of only 362 peptides or less. Moreover, all these models essentially only provided answers to cleavability or not, without any interpretability of the underlying mechanisms in a physico-chemical sense. You et al. analyzed an extended data set of substrates and found that hydrophobicity and size

The Supplementary Material referred to in this article can be found online at http://www.interscience.wiley.com/jpages/0887-3585/suppmat/

Grant sponsor: Swedish VR; Grant number: 04X-05957; Grant sponsors: Knut and Alice Wallenberg Foundation; Swedish Foundation.

*Correspondence to: Jan Komorowski, The Linnaeus Centre for Bioinformatics, Uppsala University, S-75124 Uppsala, Sweden.
E-mail: jan.komorowski@lcb.uu.se

Received 2 November 2006; Revised 5 December 2006; Accepted 2 January 2007

Published online 10 April 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21415

TABLE I. Discretization of 20 Natural Amino Acids. z-Scales Roughly Represent Hydrophobicity, Steric Properties, Polarizability (z_1, z_2, z_3), Polarity, and Electronic Effects (z_4, z_5) of Amino Acids[39]

Amino acids	Original z-scales					Discretized ternary z*-scales				
	z_1	z_2	z_3	z_4	z_5	z_1^*	z_2^*	z_3^*	z_4^*	z_5^*
Ala	0.24	−2.32	0.60	−0.14	1.30	0	−1	0	0	1
Arg	3.52	2.50	−3.50	1.99	−0.17	1	0	−1	1	0
Asn	3.05	1.62	1.04	−1.15	1.61	1	0	1	−1	1
Asp	3.98	0.93	1.93	−2.46	0.75	1	0	1	−1	1
Cys	0.84	−1.67	3.71	0.18	−2.65	1	−1	1	0	−1
Gln	1.75	0.50	−1.44	−1.34	0.66	1	0	−1	−1	0
Glu	3.11	0.26	−0.11	−3.04	−0.25	1	0	0	−1	0
Gly	2.05	−4.06	0.36	−0.82	−0.38	1	−1	0	0	0
His	2.47	1.95	0.26	3.90	0.09	1	0	0	1	0
Ile	−3.89	−1.73	−1.71	−0.84	0.26	−1	−1	−1	0	0
Leu	−4.28	−1.30	−1.49	−0.72	0.84	−1	−1	−1	0	1
Lys	2.29	0.89	−2.49	1.49	0.31	1	0	−1	1	0
Met	−2.85	−0.22	0.47	1.94	−0.98	0	0	0	1	−1
Phe	−4.22	1.94	1.06	0.54	−0.62	−1	0	1	1	−1
Pro	−1.66	0.27	1.84	0.70	2.00	0	0	1	1	1
Ser	2.39	−1.07	1.15	−1.39	0.67	1	−1	1	−1	0
Thr	0.75	−2.18	−1.12	−1.46	−0.40	0	−1	−1	−1	0
Trp	−4.36	3.94	0.59	3.44	−1.59	−1	1	0	1	−1
Tyr	−2.54	2.44	0.43	0.04	−1.47	0	0	0	0	−1
Val	−2.59	−2.64	−1.54	−0.85	−0.02	0	−1	−1	0	0

of substrate amino acids are important variables in HIV-1 protease recognition of cleavage sites.[18] However, the question about the molecular nature of the HIV-1 protease specificity remains generally unanswered and is, indeed, of great scientific interest and importance. Using an extensive data set, collected from the available experimental data from 16 years of HIV research, we created a general rule model for the HIV-1 protease specificity. We show that by using such a broad approach we can characterize in detail the complex patterns of physico-chemical specificity determinants for HIV-1 protease substrates. Our study provides novel insights into the molecular recognition mechanisms for HIV-1 protease interactome from a broad perspective.

MATERIALS AND METHODS

Data

We combined essentially all data from 16 years of HIV-1 protease and its substrates interactions research, between 1990 and 2005, into a single data set. This data set comprised 1625 substrates (374 cleavable and 1251 noncleavable) that had been experimentally tested for cleavage by the wild-type HIV-1 protease.[13,17,19–38] (Supplementary Table I).

Description of Substrates

Sandberg et al. applied principal component analysis on 26 physico-chemical properties of 87 coding and non-coding amino acids to derive their five principal components, the so-called z-scales.[39] These z-scales, roughly represent hydrophobicity, steric properties, polarizability (z_1, z_2, z_3), polarity, and electronic effects (z_4, z_5) of amino acids and are independent from each other. Each amino acid of the octapeptide sequence can thus be represented as a vector in a 5-dimensional physico-chemical property space.

We then discretized the continuous z-scales dividing each of them into three intervals. Firstly, we sorted 87 amino acids by the values of their corresponding z-scales (separately for each z_i-scale, $i = 1,\ldots,5$). We next divided each z_i-scale into three intervals using cuts so that each interval contained exactly 29 amino acids. Subsequently, these intervals were labeled −1, 0, and +1. Thus, the first 29 amino acids with the smallest corresponding z-scale values obtained label −1, the next 29 amino acids received 0 and the remaining 29 amino acids with the largest values of the corresponding z-scale were given label +1. The discretized z_i scales were named z_i^*, correspondingly. Thus, every amino acid was described by a five-ternary vector, which roughly group 87 natural and artificial amino acids according to their principal physico-chemical properties. The original and discretized z*-scales for the 20 naturally occurring amino acids are shown in Table I. Two amino acid pairs (Asp-Asn and Arg-Lys) had the same description after the discretization of their z-scales because of the similarity of their physico-chemical properties. Since sequences of six substrate pairs differed only in one amino acid of either Asp-Asn or Arg-Lys, and all these substrates were cleavable by HIV-1 protease, we removed one substrate from each pair as they were indiscernible.

Finally, we described HIV-1 protease substrate sequences spanning from P_4 to P_4' position by the five-ternary z*-scales, which yielded 40 attributes per substrate. Each sequence was given a label, +1 or −1 that stated

whether the substrate was, respectively, cleavable or not. This process resulted in a table with substrates as rows, their physico-chemical properties as columns and entries 1 or −1 as the decision attribute.

Rough Sets

Rough set theory constitutes a mathematical framework for inducing minimal decision IF-THEN rules from labeled examples.[40–42] A decision rule consists of a conjunction of attribute values and a decision value. We used this framework, implemented in the Rosetta[43] system (http://rosetta.lcb.uu.se), for learning IF-THEN rules associating the physico-chemical properties of substrate amino acids with the HIV-1 protease cleavability. Since computing minimal sets of attributes is computationally very expensive, we used a genetic algorithm to search for approximate solutions, called approximate reducts. IF-THEN rules were then constructed on these reducts to obtain a link between minimal combination of amino acids properties and HIV-1 protease cleavability. All such rules together comprise a rule-based model with a capacity of predicting the cleavability of unknown sequences by HIV-1 protease. Such a rule-based model is inherently descriptive and the patterns appearing in the labeled data can be easily interpreted.

The rough set-based method was thus used to generate a model from the 1619 training examples. Since the obtained model consisted of over 136,000 highly specific rules, we applied a rule-tuning algorithm,[44] (i.e., the group generalization method implemented in the Rosetta[43] system), to simplify the rules. Rule group generalization works with groups of similar rules. The assignment of rules to their groups is based on the assumption that merging rules with the same decision and overlapping sets of attributes prevents a high drop in accuracy of the generated rule in comparison to the original ones. Application of the rule-tuning method allowed a significant reduction of the size of the model down to 408 general rules without a significant drop in accuracy and AUC values (<0.01) (See Validation section below). The importance of the rules was evaluated by rule support, coverage, and rule accuracy parameters. Rule support is the number of objects in the decision system having the property described by the conditional part of the rule (left hand side, LHS). In contrast, rule coverage shows the proportion of the objects in the training set that are identifiable by this rule, while accuracy reflects a ratio between correctly classified objects identifiable by the rule to the total number of objects identifiable by the rule.

Validation

The validity of the model was determined using a double cross-validation (CV) procedure.[45] First, the data set of substrates was randomly divided into five equally sized subsets D_i, $i = 1,2,\ldots,5$. We then generated five new data sets of substrates (N_i) removing once one of the D_i subsets from the original set of substrates in each new set. Thus, the N_1 data set comprised D_2, D_3, D_4, and D_5 subsets; the N_2 data set comprised D_1, $D_3,\ldots,D_5$ and so on. We then applied an internal 10-fold CV procedure. Each of the $N_1, \ldots, N_5$ data sets of substrates was separately and randomly divided into 10 equally sized subsets. A rule classifier was induced as described above from the nine subsets (the training set) and used to classify the substrates in the remaining tenth subset (the test set). This procedure was repeated 10 times, so that each substrate was in the test set once and in the training set nine times.

We then performed an external fivefold CV procedure. A rule classifier was induced as described previously from each of the $N_1,\ldots,N_5$ data sets separately and then used to classify the substrates in the corresponding D_i subset not included in the particular N_i data set. For example, rules induced from the N_1 data set were used to predict substrates included in the D_1 subset.

We then evaluated models performance using prediction accuracy and area under the ROC curve (AUC) mean values. The accuracy mean is an average fraction of correctly predicted objects computed for n blocks during the n-fold CV test. The Receiver Operating Characteristics (ROC) curve estimates the threshold independent performance of the classifier by plotting *sensitivity* against (*1-specificity*). *Sensitivity* is a ratio between true positive predictions and the sum of true positive and false negative predictions, whereas (*1-specificity*) is a ratio between false positive predictions and the sum of true negative and false positive predictions. For a binary classifier, an AUC of 1.0 indicates that the discriminatory power is perfect while an AUC of 0.5 denotes that the classification of objects is random. The AUC mean value is an average AUC for the models induced by the n blocks during n-fold CV test.

Finally, we performed a randomization validation test, where the decision attribute was repeatedly and randomly permutated, yielding new data set samples with replacements from the original data set. We then performed external fivefold CV for 100 new data set samples as described above and computed AUC and accuracy mean values in each case. We then counted the fraction of times when AUC and accuracy mean values for the permutated data sets were larger than AUC and accuracy mean values obtained for the original data set. This fraction may be interpreted as a *P*-value, i.e. the probability that the relationship found in the original data set is obtained by chance.

RESULTS AND DISCUSSION

Application of rough sets approach on the largest compiled substrate set resulted in a statistically valid rule-based model for the HIV-1 protease specificity (Table II). We used accuracy mean and area under the ROC curve values to evaluate model performance. First, internal 10-fold CV for the models $N_1,\ldots,N_5$, constructed on the 4/5 of the initial data set, showed very good results with high prediction accuracy mean 93 ± 2% and high AUC

TABLE II. External and Internal CV Results for Models $N_1,\ldots,N_5$

Models	Internal 10-fold CV		Prediction of D data set	External 5-fold CV	
	Accuracy mean (SD)	AUC mean (SD)		Accuracy mean	AUC mean
Model N_1	0.93 (0.03)	0.94 (0.03)	D_1	0.92	0.92
Model N_2	0.93 (0.02)	0.95 (0.03)	D_2	0.94	0.93
Model N_3	0.93 (0.02)	0.95 (0.03)	D_3	0.94	0.94
Model N_4	0.93 (0.03)	0.94 (0.03)	D_4	0.94	0.96
Model N_5	0.93 (0.02)	0.95 (0.03)	D_5	0.93	0.94
Average	**0.93 (0.02)**	**0.95 (0.03)**		**0.93**	**0.94**

SD Denotes Standard Deviation of Accuracy or AUC. D_1, …, D_5 Denotes the External Data Set for the Corresponding N_1, …, N_5 Model.

TABLE III. Nine Rules in the Model for Cleavable Substrates With the Largest Support

No.	Rules	Rule support	Rule accuracy (%)	LHS coverage	RHS coverage	Rule length
1	If $P_2z_1^*(0)$ ^ $P_2z_4^*(0)$ ^ $P_2'z_3^*(0)$ ^ $P_3'z_3^*(0)$ ^ $P_4'z_1^*(0)$ then cleavage	86	100	0.05	0.23	5
2	If $P_2z_1^*(0)$ ^ $P_2z_2^*(-1)$ ^ $P_2'z_3^*(0)$ ^ $P_3'z_3^*(0)$ ^ $P_4'z_1^*(0)$ then cleavage	86	100	0.05	0.23	5
3	If $P_4z_1^*(0)$ ^ $P_2z_1^*(0)$ ^ $P_2z_3^*(-1)$ ^ $P_3'z_3^*(0)$ ^ $P_4'z_1^*(0)$ then cleavage	75	100	0.05	0.20	5
4	If $P_4z_1^*(0)$ ^ $P_2z_1^*(0)$ ^ $P_1z_2^*(-1)$ ^ $P_1'z_4^*(0)$ ^ $P_4'z_1^*(0)$ then cleavage	74	100	0.05	0.20	5
5	If $P_4z_1^*(0)$ ^ $P_2z_4^*(0)$ ^ $P_1'z_5^*(1)$ ^ $P_3'z_3^*(0)$ ^ $P_4'z_1^*(0)$ then cleavage	71	100	0.04	0.19	5
6	If $P_2'z_3^*(0)$ ^ $P_3'z_1^*(0)$ ^ $P_3'z_5^*(-1)$ ^ $P_4'z_1^*(0)$ ^ $P_4'z_4^*(-1)$ then cleavage	70	100	0.04	0.19	5
7	If $P_2'z_4^*(0)$ ^ $P_3'z_1^*(0)$ ^ $P_3'z_3^*(0)$ ^ $P_3'z_4^*(1)$ ^ $P_4'z_2^*(-1)$ then cleavage	69	100	0.04	0.19	5
8	If $P_1'z_1^*(-1$ or $0)$ ^ $P_3'z_1^*(0)$ ^ $P_3'z_5^*(-1)$ ^ $P_4'z_2^*(-1)$ ^ $P_4'z_4^*(-1)$ then cleavage	68	100	0.04	0.18	5
9	If $P_2'z_1^*(0)$ ^ $P_2'z_3^*(0)$ ^ $P_3'z_1^*(0)$ ^ $P_3'z_5^*(-1)$ ^ $P_4'z_2^*(-1)$ then cleavage	68	100	0.04	0.18	5

LHS and RHS Coverage Denote Rule Left Hand Side (Conditional *if* Part) and Rule Right Hand Side (Decision *then* Part) Coverage, Respectively.

mean 0.95 ± 0.03. We then performed external fivefold CV procedure for the original data set. Again, the models demonstrated high prediction accuracy with respect to new data (accuracy mean 93%) and high classification quality (AUC mean 0.94). Fivefold CV can in this case be considered as an external validation test for each of the models $N_1,\ldots,N_5$, individually. Results of the randomization validation test demonstrated further that it is very unlikely to obtain valid models based on random data (accuracy mean 76 ± 2%, AUC mean 0.50 ± 0.03, P-value < 0.01).

We analyzed the rule model and identified 218 rules for cleavable and 190 rules for noncleavable substrates (Tables III and IV, see Supplementary Tables II and III for a full list of rules). The analysis of the rules demonstrates that amino acids in at least three substrate positions define a pattern necessary for processing by HIV-1 protease, since no less than three physico-chemical properties for three substrate positions are present in each rule for cleavage (Table III, Supplementary Table II). This novel finding extends a traditional opinion about the major importance of amino acids in positions P_2,P_1,P_1',P_2' for HIV-1 substrates.[3,46–49] The relation is much more complex than previously described and many rules are necessary to model all the interactions between substrate amino acid properties and cleavability. Instead, each rule covers a set of substrate properties and associates them with cleavability. Altogether, these rules constitute a general model of the HIV-1 protease specificity. The further analysis of the rules shows the most necessary physico-chemical properties of amino acids for substrate cleavability by HIV-1 protease, i.e. hydrophobicity for P_1' and P_2 positions ($P_1'z_1^*$ and $P_2z_1^*$ attributes are present in 61 and 59 rules respectively), polarity for P_4' position ($P_4'z_4^*$ is present in 53 rules), polarizability for P_2' and P_4 positions ($P_2'z_3^*$ and $P_4z_3^*$ are found in 52 and 50 rules respectively) and electronic effects of amino acids for P_1 position ($P_1z_5^*$ is found in 50 rules) (Supplementary Table II). However, the properties of the amino acids in the above positions alone do not define HIV-1 protease specificity. Rather, it is their combination with other amino acids physico-chemical properties captured

TABLE IV. Ten Rules in the Model for Non-cleavable Substrates With the Largest Support

No.	Rules	Rule support	Rule accuracy (%)	LHS coverage	RHS coverage	Rule length
1	If $P_4z_3^*(-1 \text{ or } 0)$ ^ $P_1z_5^*(0)$ ^ $P_1'z_4^*(-1 \text{ or } 1)$ then no cleavage	379	100	0.23	0.30	3
2	If $P_1z_1^*(0 \text{ or } 1)$ ^ $P_1z_3^*(-1)$ ^ $P_3'z_5^*(0 \text{ or } 1)$ then no cleavage	371	100	0.23	0.30	3
3	If $P_1z_5^*(0)$ ^ $P_1'z_5^*(0 \text{ or } 1)$ ^ $P_4'z_1^*(1)$ then no cleavage	366	100	0.23	0.29	3
4	If $P_1z_3^*(-1)$ ^ $P_1z_5^*(0)$ ^ $P_1'z_1^*(0 \text{ or } 1)$ ^ $P_4'z_5^*(0 \text{ or } 1)$ then no cleavage	360	100	0.22	0.29	4
5	If $P_1z_3^*(-1)$ ^ $P_1z_5^*(0)$ ^ $P_3'z_3^*(-1 \text{ or } 1)$ then no cleavage	357	100	0.22	0.29	3
6	If $P_1z_5^*(0)$ ^ $P_1'z_1^*(1)$ ^ $P_3'z_5^*(0 \text{ or } 1)$ then no cleavage	351	100	0.22	0.28	3
7	If $P_3z_2^*(-1 \text{ or } 0)$ ^ $P_2z_1^*(-1 \text{ or } 1)$ ^ $P_1z_3^*(-1)$ ^ $P_1z_5^*(0)$ then no cleavage	344	100	0.21	0.27	4
8	If $P_1z_3^*(-1)$ ^ $P_1z_5^*(0)$ ^ $P_3'z_1^*(-1 \text{ or } 1)$ then no cleavage	343	100	0.21	0.27	3
9	If $P_1z_3^*(-1)$ ^ $P_1z_5^*(0)$ ^ $P_2'z_3^*(-1 \text{ or } 1)$ then no cleavage	340	100	0.21	0.27	3
10	If $P_4z_1^*(-1 \text{ or } 1)$ ^ $P_1z_3^*(-1)$ ^ $P_1z_5^*(0)$ then no cleavage	336	100	0.21	0.27	3

LHS and RHS Coverage Denotes Rule Left Hand Side (Conditional *if* Part) and Rule Right Hand Side (Decision *then* Part) Coverage, Respectively.

by our rules that eventually allow the enzyme to process a scissile bond of the substrate. It has been previously shown that HIV-1 protease prefers large hydrophobic residues at P_1 and P_1' positions, smaller hydrophobic residues at P_2 and can accommodate a variety of residues at P_3 and P_3' positions.[3,47] Our findings also demonstrate the major importance of hydrophobicity (z_1^*) for substrate amino acids in P_1' and P_2 positions as well as they capture new physico-chemical features and their combinations preferential for HIV-1 protease cleavage.

Our analysis of the rules for non-cleavage indicated that the most frequently found attributes are $P_1z_5^*$ (present in 63 rules), $P_1z_1^*$ (present in 50 rules), $P_2z_1^*$ (present in 44 rules) and $P_1'z_1^*$ (present in 44 rules) (Supplementary Table III). This finding suggests that amino acids in P_1 position with properties z_1^* or z_5^* equal to 0 or 1 would likely prevent substrate cleavage by HIV-1 protease. Similarly, a substrate is likely to be non-cleavable by the enzyme if it has an amino acid with the property $z_1^* = -1$ or $z_1^* = 1$ in position P_2 or $z_1^* = 1$ in position P_1'. These findings align well with the previously discovered rules for amino acid residues not tolerated in the HIV-1 protease substrates.[47,49] For example, Lys is not tolerated from positions P_2 through P_2' and aromatic amino acids are not preferred in positions P_2 or P_2' positions.[49] These conditions are readily captured by our general model since $z_1^* = 1$ for Lys and $z_1^* = -1$ for aromatic amino acids in P_2-P_2' positions would satisfy the conditional part of the rules for non-cleavage. Additionally, β-branched amino acids, Ser and Gly would most likely prevent cleavage by the protease if they were present in position P_1.[47,49] In our model substrates with β-branched amino acids, Ser and Gly in position P_1 would then be recognized by the rules for non-cleavage since these amino acids have property $z_5^* = 0$. However, the above statements are true if amino acids in P_2, P_1, and P_1' positions coexist with particular amino acids in other positions according to the rules (Table IV and Supplementary Table III).

Our model may be also used to generate substrate sequences that are cleavable. Indeed, it is straightforward to convert the physico-chemical ternary descriptors into amino acids. For example, rule 2 in Table III states that if there are amino acids with properties $z_1^* = 0$ and $z_2^* = -1$ in position P_2, and amino acids with property $z_3^* = 0$ in position P_2', and amino acids with property $z_3^* = 0$ in position P_3', and amino acids with property $z_1^* = 0$ in position P_4' then the substrate is cleavable. This rule can be interpreted in the language of amino acids, as follows: if the position P_2 has Ala, Thr, or Val (the only amino acids with $z_1^* = 0$ and $z_2^* = -1$ at the same time, see Table I), and P_2' has Ala, Gly, Met, Glu, or His, and P_3' has Ala, Glu, Gly, His, Met, Trp, or Tyr, and P_4' has Ala, Met, Pro, Thr, or Val then the substrate is cleavable [rule 2 in Fig. 1(a)]. In the same way, we translated other rules into matching amino acids to grasp their biological meaning (Fig. 1). The comprehensive maps of HIV-1 protease specificity are shown in Supplementary Tables IV and V and provide a general biological interpretation of our model. Although a substrate satisfying the requirements of rule 2 in Table III may have any amino acid in position P_1', the alignment of cleavable substrates shows that some amino acids are never pres-

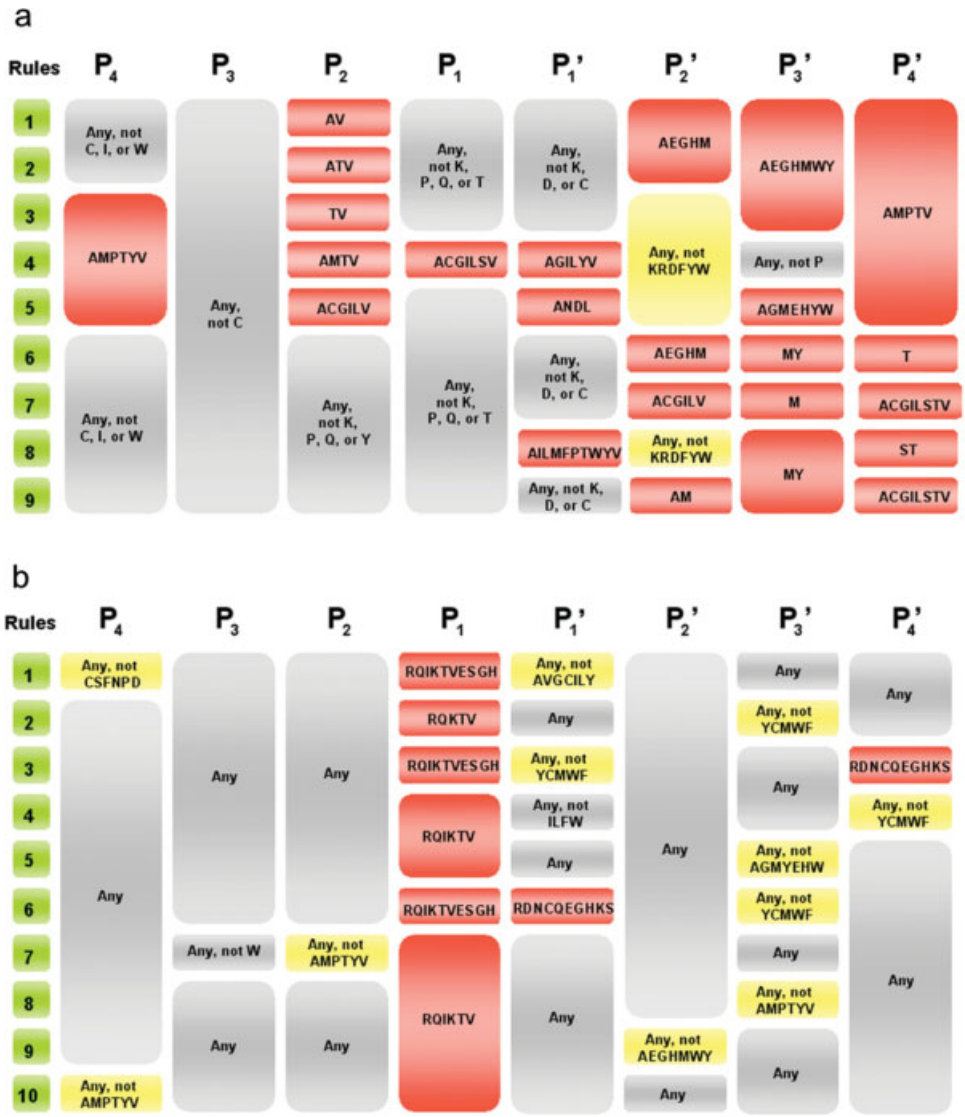

Fig. 1. Representation of ten rules with the largest support in terms of amino acids for cleavable substrates (**a**) and for non-cleavable substrates (**b**). Each column shows the set of amino acids in a substrate P_i position that match a particular rule. Red color boxes indicate that 1 to 10 amino acids satisfy the rule conditions. Yellow color boxes denote that 11–15 amino acids meet rule requirements. Grey color boxes indicate that 16–20 natural coding amino acids may occur in P_i position. The amino acids residues are presented by the standard one-letter code.

TABLE V. Amino Acids Absent in the Set of Substrates Cleavable by HIV-1 Protease

Position	Amino acids
P_4	C, I, W
P_3	C
P_2	K, P, Q, Y
P_1	K, P, Q, T
P_1'	C, D, K
P_2'	D, F, K, R, W, Y
P_3'	P
P_4'	W, Y

ent in the particular positions (Table V). For example, Lys never occurs in any of the positions spanning from P_2 to P_2' in known cleavable sequences. Accordingly, we excluded such amino acids from the interpretation of rules for cleavable sequences [Fig. 1(a)]. On the contrary, all the 20 coding amino acids are present at least once in all positions within the non-cleavable substrate set.

We then analyzed P_1-P_1' amino acid pairs in cleavable octapeptides included in the study and found that there are 92 different P_1-P_1' amino acid combinations within the cleavable substrate set. Our data shows that only 43 P_1-P_1' pairs match substrate class "i" or "ii" according to the traditional HIV-1 protease substrate classification,[3,48,49] whereas the majority of P_1-P_1' pairs (53%) fall into class "iii" that contains hydrophilic amino acid residues either in P_1 or in P_1'. Moreover, substrates with Ser-Arg, Ser-Gln, Ser-Glu, Ser-Gly, Asn-Thr, or Asn-Gly amino acids in the P_1-P_1' positions, respectively, are also

processed by the enzyme. Noteworthy, 92% of P_1-P_1' amino acid combinations, found within the cleavable substrate set, are present within the non-cleavable substrate set. Therefore, the determination of HIV-1 protease specificity based on the amino acids in the P_1-P_1' positions[3,48,49] covers only a small part of the substrate cleavage requirements for this important enzyme clearly supporting our findings. One possible explanation for the prevailing classification may be the hydrophobic nature of P_1-P_1' amino acids in a large fraction of natural cleavage sites in Gag and Gag-Pol polyproteins, which may be required to ensure efficient processing of the polypeptide chain by the enzyme. Although cleavage sites falling within class "iii" are less frequently observed in nature, we anticipate that this is due to posttranslational *in vivo* modification of the hydrophilic amino acid functional groups forming these sites, e.g. phosphorylation or glycosylation of Ser, Thr or Asp. Such modification eventually prevents substrate hydrolysis, which otherwise would occur.[25] This could provide a mechanism for cleavage-protection that might play a regulatory role in the maturation of viral particles. Alternatively, some cleavage sites belonging to class iii) might be hidden by the 3D structure of the polyproteins and, therefore, inaccessible for HIV-1 protease.

CONCLUSIONS

In the present study, we developed a new approach for the HIV-1 protease interactome studies. We demonstrated that HIV-1 protease specificity is much more complex than previously anticipated, and therefore cannot be defined based solely on substrate amino acids in positions P_1 and P_1' or any other single position. Using rough set-based rule modeling we provided a general approach to analyzing HIV-1 protease specificity and showed that a combination of at least three particular amino acids is needed for a cleavage event to occur. The rule-based model not only allows accurate prediction of the substrate cleavage but it also provides the most detailed characterization of HIV-1 protease specificity made to date. The model allows interpreting the complex physico-chemical features of cleavage in a simple visual format. Our study clearly demonstrates that by combining and analyzing large HIV-1 protease interactome data it is possible to discover novel and general patterns of physico-chemical determinants for protease-peptide interactions. We foresee that this method will provide a platform to test variety of new hypotheses in viral proteomics field and encourage other studies toward the generalized use of interactome data.

ACKNOWLEDGMENTS

The authors thank Dr. Torgeir R. Hvidsten, Ewa Makosa, Marcin Kierczak and Lukasz Ligowski for the help with Rosetta and technical help in running computations. The authors thank Dr. Peteris Prusis for a critical review and useful comments in preparation of the manuscript.

REFERENCES

1. Coffin JM. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science 1995; 267:483–489.
2. Perelson AS, Neumann AU, Markowitz M, Leonard JM, Ho DD. HIV-1 dynamics in vivo: virion clearance rate, infected cell lifespan, and viral generation time. Science 1996;271:1582–1586.
3. Beck ZQ, Morris GM, Elder JH. Defining HIV-1 protease substrate selectivity. Curr Drug Targets Infect Disord 2002;2:37–50.
4. Dauber DS, Ziermann R, Parkin N, Maly DJ, Marhus S, Harris JL, Ellman JA, Petropoulos C, Craik CS. Altered substrate specificity of drug-resistant human immunodeficiency virus type 1 protease. J Virol 2002;76:1359–1368.
5. Randolph JT, DeGoey DA. Peptidomimetic inhibitors of HIV protease. Curr Top Med Chem 2004;4:1079–1095.
6. Wlodawer A, Gustchina A. Structural and biochemical studies of retroviral proteases. Biochim Biophys Acta 2000;1477:16–34.
7. Cai Y-D, Chou K-C. Using neural network for prediction of HIV protease cleavage sites in proteins. J Protein Chem 1998;17: 607–615.
8. Thomson R, Hodgman TC, Yang ZR, Doyle AK. Characterising proteolytic cleavage site activity using bio-basis function neural network. Bioinformatics 2003;19:1741–1747.
9. Yang ZR, Thomson R. Bio-basis function neural network for prediction protease cleavage sites in proteins. IEEE Trans Neural Netw 2005;16:263–274.
10. Narayanan A, Wu XK, Yang ZR. Mining viral protease data. to extract cleavage knowledge. Bioinformatics 2002;18:5–13.
11. Yang ZR, Thomson R, Hodgman TC, Dry J, Doyle AK, Narayanan A, Wu X. Extracting decision rules from protein sequences using genetic programming methods. Biosystems 2003;72: 159–176.
12. Yang ZR, Chou K-C. Bio-support vector machines for computational proteomics. Bioinformatics 2004;20:735–741.
13. Poorman RA, Tomasselli AG, Heinrikson RL, Kezdy FJ. A cumulative specificity model for proteases from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. J Biol Chem 1991;22:14554–14561.
14. Chou JJ. A formulation for correlating properties of peptides and its application to predicting human immunodeficiency virus protease-cleavable sites in proteins. Biopolymers 1993;33:1405–1414.
15. Chou JJ. Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. J Protein Chem 1993; 12:291–302.
16. Chou K-C. A vectorised sequence-coupling model for predicting HIV protease cleavage sites in proteins. J Biol Chem 1993;268: 16938–16948.
17. Chou K-C., Tomasselli AG, Reardon IM, Hendrikson RL. Predicting human immunodeficiency virus protease cleavage sites in proteins by a discriminant function method. Proteins 1996;24: 51–72.
18. You L, Garwicz D, Rögnvaldsson T. Comprehensive bioinformatic analysis of the specificity of human immunodeficiency virus type 1 protease. J Virol 2005;79:12477–12486.
19. Ridky TW, Kikonyogo A, Leis J, Gulnik S, Copeland T, Erickson J, Wlodawer A, Kurinov I, Harrison RW, Weber IT. Drug-resistant HIV-1 proteases identify enzyme residues important for substrate selection and catalytic rate. Biochemistry 1998;37:13835–13845.
20. Cameron CE, Grinde B, Jacques P, Jentoft J, Leis J, Wlodawer A, Weber IT. Comparison of the substrate-binding pockets of the Rous sarcoma virus and human immunodeficiency virus type 1 proteases. J Biol Chem 1993;268:11711–11720.
21. Ridky TW, Cameron CE, Cameron J, Leis J, Copeland T, Wlodawer A, Weber IT, Harrison RW. Human immunodeficiency virus type 1 protease substrate specificity is limited by interactions between substrate amino acids bound in adjacent enzyme subsites. J Biol Chem 1996;271:4709–4717.
22. Tozser J, Gustchina A, Weber IT, Blaha I, Wondrak EM, Oroszlan S. Studies on the role of the S4 substrate binding site of HIV proteinases. FEBS Lett 1991;279:356–360.
23. Cameron CE, Grinde B, Jentoft J, Leis J, Weber IT, Copeland TD, Wlodawer A. Mechanism of inhibition of the retroviral protease by a Rous sarcoma virus peptide substrate representing

the cleavage site between the gag p2 and p10 proteins. J Biol Chem 1992;267:23735–23741.
24. Tozser J, Bagossi P, Weber IT, Copeland TD, Oroszlan S. Comparative studies on the substrate specificity of avian myeloblastosis virus proteinase and lentiviral proteinases. J Biol Chem 1996;271:6781–6788.
25. Tozser J, Bagossi P, Boross P, Louis JM, Majerova E, Oroszlan S, Copeland TD. Effect of serine and tyrosine phosphorylation on retroviral proteinase substrates. Eur J Biochem 1999;265: 423–429.
26. Boross P, Bagossi P, Copeland TD, Oroszlan S, Louis JM, Tozser J. Effect of substrate residues on the P2′ preference of retroviral proteinases. Eur J Biochem 1999;264:921–929.
27. Louis JM, Oroszlan S, Tozser J. Stabilization from autoproteolysis and kinetic characterization of the human T-cell leukemia virus type 1 proteinase. J Biol Chem 1999;274:6660–6666.
28. Tozser J, Weber IT, Gustchina A, Blaha I, Copeland TD, Louis JM, Oroszlan S. Kinetic and modeling studies of S3-S3′ subsites of HIV proteinases. Biochemistry 1992;31:4793–4800.
29. Tozser J, Bagossi P, Weber IT, Louis JM, Copeland TD, Oroszlan S. Studies on the symmetry and sequence context dependence of the HIV-1 proteinase specificity. J Biol Chem 1997;272:16807–16814.
30. Tozser J, Blaha I, Copeland TD, Wondrak EM, Oroszlan S. Comparison of the HIV-1 and HIV-2 proteinases using oligopeptide substrates representing cleavage sites in Gag and Gag-Pol polyproteins. FEBS Lett 1991;281:77–80.
31. Kadas J, Weber IT, Bagossi P, Miklossy G, Boross P, Oroszlan S, Tozser J. Alteration of the specificity of human T-cell leukemia virus type-1 protease. J Biol Chem 2004;279:27148–27157.
32. Feher A, Weber IT, Bagossi P, Boross P, Mahalingam B, Louis JM, Copeland TD, Torshin IY, Harrison RW, Tozser J. Effect of sequence polymorphism and drug resistant mutations at two HIV-1 Gag cleavage sites on their proteolytic susceptibility. Eur J Biochem 2002;269:4114–4120.
33. Tozser J, Zahuczky G, Bagossi P, Louis JM, Copeland TD, Oroszlan S, Harrison RW, Weber IT. Comparison of the substrate specificity of the human T-cell leukemia virus and human immunodeficiency virus proteinases. Eur J Biochem 2000;267: 6287–6295.
34. Partin K, Krausslich H-G, Ehrlich L, Wimmer E, Carter C. Mutational analysis of a native substrate of the human immunodeficiency virus type 1 proteinase. J Virol 1990;64:3938–3947.
35. Beck ZQ, Lin Y-C, Elder JH. Molecular basis for the relative substrate specificity of human immunodeficiency virus type 1 and feline immunodeficiency virus proteases. J Virol 2001;75: 9458–9469.
36. Beck ZQ, Hervio L, Dawson PE, Elder JH, Madison EL. Identification of efficiently cleaved substrates for HIV-1 protease using a phage display library and use in inhibitor development. Virology 2000;274:391–401.
37. Shoeman RL, Honer B, Stoller TJ, Kesselmeier C, Miedel MC, Traub P, Graves MC. Human immunodeficiency virus type 1 protease cleaves the inermediate filament proteins vimentin, desmin, and glial fibrillary acidic protein. Proc Natl Acad Sci USA 1990;87:6336–6340.
38. Tomasselli AG, Sarcich JL, Barrett LJ, Reardon IM, Howe WJ, Evans DB, Sharma SK, Heinrikson RL. Human immunodeficiency virus type-1 reverse transcriptase and ribonuclease H as substrates of the viral protease. Protein Sci 1993;2:2167–2176.
39. Sandberg M, Eriksson L, Jonsson J, Sjöström M, Wold S. New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 1998;4:2481–2491.
40. Pawlak Z. Rough sets. Theory Decis Libr 1991;9:1–4.
41. Komorowski J. Rough sets: A tutorial. In: Pal SK, Skowron A, editors. Rough-fuzzy hybridization: a new trend in decision making. Singapore: Springer Verlag; 1999. pp 3–98.
42. Skowron A, Komorowski J, Pawlak Z, Polkowski L. Rough sets perspective on data and knowledge. In: Kløsgen W, Zytkow J, editors. Handbook of data mining and knowledge discovery. New York: Oxford University Press; 2002. pp 134–149.
43. Komorowski J, Øhrn A, Skowron A. ROSETTA Rough sets. In: Kløsgen W, Zytkow J, editors. Handbook of data mining and knowledge discovery. New York: Oxford University Press; 2002. pp 554–559.
44. Makosa E. Rule tuning (masters' thesis). The Linnaeus Centre for Bioinformatics, Uppsala University, Uppsala, Sweden; 2005.
45. Freyhult E, Prusis P, Lapinsh M, Wikberg JES, Moulton V, Gustafsson MG. Unbiased descriptor and parameter selection confirms the potential of proteochemometric modelling. BMC Bioinformatics 2005;6:50.
46. Bagossi P, Sperka T, Feher A, Kadas J, Zahuczky G, Miklossy G, Boross P, Tozser J. Amino acid preferences for a critical substrate binding subsite of retroviral proteases in type 1 cleavage sites. J Virol 2005;79:4213–4218.
47. Louis JM, Weber IT, Tozser J, Clore GM, Gronenborn AM. HIV-1 protease: maturation, enzyme specificity, and drug resistance. Adv Pharmacol 2000;49:111–146.
48. Pettit SC, Simsic J, Loeb DD, Everitt L, Hutchison CA 3rd, Swanstrom R. Analysis of retroviral protease cleavage sites reveals two types of cleavage sites and the structural requirements of the P1 amino acid. J Biol Chem 1991;266:14539–14547.
49. Tomasselli AG, Heinrikson RL. Specificity of retroviral proteases: analysis of viral and nonviral protein substrates. Methods Enzymol 1994;241:279–301.