DiVA
http://www.diva-portal.org

Postprint

This is the accepted version of a paper published in *Journal of Virology*. This paper has been peer-reviewed but does not include the final publisher proof-corrections or journal pagination.

Citation for the original published paper (version of record):

You, L., Garwicz, D., Rögnvaldsson, T. (2005)
Comprehensive Bioinformatic Analysis of the Specificity of Human Immunodeficiency Virus Type 1 Protease.
Journal of Virology, 79(19): 12477-12486
http://dx.doi.org/10.1128/JVI.79.19.12477-12486.2005

Access to the published version may require subscription.

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:hh:diva-268

JOURNAL OF VIROLOGY, Oct. 2005, p. 12477–12486 Vol. 79, No. 19
0022-538X/05/$08.00+0 doi:10.1128/JVI.79.19.12477–12486.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

Comprehensive Bioinformatic Analysis of the Specificity of Human Immunodeficiency Virus Type 1 Protease

Liwen You,[1] Daniel Garwicz,[2] and Thorsteinn Rögnvaldsson[1]*

School of Information Science, Computer and Electrical Engineering, Halmstad University, Halmstad, Sweden,[1] and Division of Hematology and Transfusion Medicine, Department of Laboratory Medicine, Lund University, Lund, Sweden, and Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden[2]

Received 28 November 2004/Accepted 1 July 2005

Rapidly developing viral resistance to licensed human immunodeficiency virus type 1 (HIV-1) protease inhibitors is an increasing problem in the treatment of HIV-infected individuals and AIDS patients. A rational design of more effective protease inhibitors and discovery of potential biological substrates for the HIV-1 protease require accurate models for protease cleavage specificity. In this study, several popular bioinformatic machine learning methods, including support vector machines and artificial neural networks, were used to analyze the specificity of the HIV-1 protease. A new, extensive data set (746 peptides that have been experimentally tested for cleavage by the HIV-1 protease) was compiled, and the data were used to construct different classifiers that predicted whether the protease would cleave a given peptide substrate or not. The best predictor was a nonlinear predictor using two physicochemical parameters (hydrophobicity, or alternatively polarity, and size) for the amino acids, indicating that these properties are the key features recognized by the HIV-1 protease. The present in silico study provides new and important insights into the workings of the HIV-1 protease at the molecular level, supporting the recent hypothesis that the protease primarily recognizes a conformation rather than a specific amino acid sequence. Furthermore, we demonstrate that the presence of 1 to 2 lysine residues near the cleavage site of octameric peptide substrates seems to prevent cleavage efficiently, suggesting that this positively charged amino acid plays an important role in hindering the activity of the HIV-1 protease.

In less than a quarter of a century, over 20 million people have succumbed to AIDS, and at the end of 2003, an estimated 38 million people were living with a human immunodeficiency virus (HIV) infection. With an increase of almost 5 million new cases per year, more than 40 million people are likely to be infected with HIV today, with over 2 million of those afflicted being children under the age of 15 years (see the UNAIDS Report on the Global AIDS Epidemic and the AIDS Epidemic Update December 2004 from UNAIDS/WHO [48a, 48b]).

Drugs that inhibit the HIV-1 protease, so-called protease inhibitors, are an important part of AIDS therapy today (20), since the HIV-1 protease cleaves viral Gag and Gag-Pol polyproteins into structure and replication proteins that are necessary for the virus to become infectious (28). Currently licensed protease inhibitors are all peptidomimetic; they mimic a peptide that the HIV-1 protease normally cleaves but are chemically modified such that the scissile bond cannot be cleaved (21, 37). Hence, rational design of an efficient inhibitor requires a good understanding of the HIV-1 protease specificity, i.e., knowing which amino acid sequences are cleaved by the protease and which are not. This is, however, difficult since it cleaves at several different sites that have little or no sequence similarity.

A problem with the clinical use of protease inhibitors is the fact that the virus is able to develop drug-resistant strains (8, 19). This is of course due to its high mutation rate (14) but possibly also because current inhibitors do not exploit all aspects of the HIV-1 protease specificity (10, 20). A better understanding of the HIV-1 specificity is probably necessary for developing more efficient inhibitors.

The HIV-1 protease has an active site with eight subsites, denoted S_4-S_3-S_2-S_1-$S_{1'}$-$S_{2'}$-$S_{3'}$-$S_{4'}$, where eight corresponding residues can be bound, denoted P_4-P_3-P_2-P_1-$P_{1'}$-$P_{2'}$-$P_{3'}$-$P_{4'}$ (the scissile bond is located between P_1 and $P_{1'}$). There is no known algorithm (rule) yet that, given a sequence of eight amino acids, P_4-P_3-P_2-P_1-$P_{1'}$-$P_{2'}$-$P_{3'}$-$P_{4'}$ (an octamer), can tell whether it will be cleaved by the HIV-1 protease or not. Several researchers have, however, attempted to construct such an algorithm using various methods. A first attempt was the "*h*-function algorithm" by Poorman et al. (34), and Chou (15) reviewed several early attempts at predicting HIV-1 protease cleavage, including the "*h* function." Thompson et al. (43) were the first to apply artificial neural networks to the problem, with some success. Cai and Chou (11) later collected a larger data set with 362 peptides, which until now has been the standard data set used, and trained a neural network to predict cleavage or noncleavage. They achieved a 92% correct prediction rate on a separate set with 63 peptides. Narayanan et al. (31) used the same data and a decision tree to extract rules for the HIV-1 protease cleavage, with rather poor results. Cai et al. (12) used support vector machines (SVMs) and concluded that the resulting predictor was inferior to the neural network presented previously by Cai and Chou in 1998. Thomson et al.

* Corresponding author. Mailing address: School of Information Science, Computer and Electrical Engineering, Halmstad University, Box 823, SE-301 18 Halmstad, Sweden. Phone: 46 35 16 74 77. Fax: 46 35 12 03 48. E-mail: denni@ide.hh.se.

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

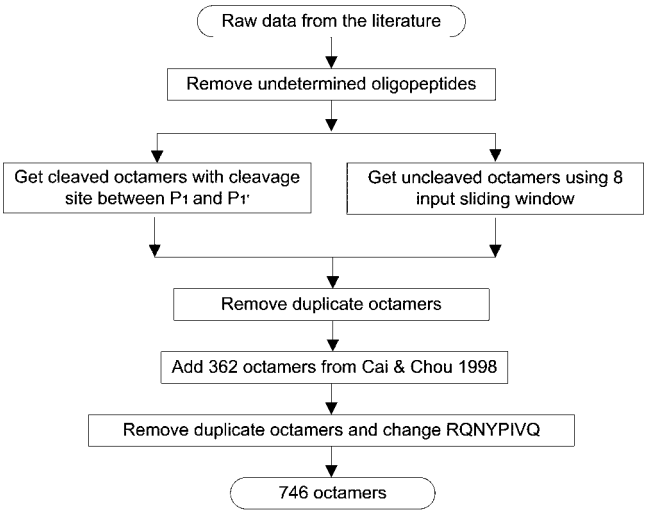

FIG. 1. Overview of the data collection procedure.

(44) recently introduced biobasis functions (BBF) for this problem, where prior knowledge is used in the form of similarity matrices, claiming that a biobasis classifier was 93% correct and better than a neural network. This idea was exploited further by Yang and Chou (51), who used biosupport functions and achieved a 91% correct classification with this approach. Yang et al. (50, 52) have also recently presented results with some sophisticated rule extraction methods based on genetic programming. However, the value of the latter approaches is questionable since they are time consuming and do not yield very accurate rules.

It should be clear from this brief review that several different algorithms have been tried on this problem over the last decade, and the conclusions have not been very consistent. It is also somewhat surprising that the same data set has always been used (the 362 peptides collected by Cai and Chou in 1998). A lot of sequence data have been published since 1998, and 362 peptides are few compared with the possible complexity of the task. In fact, we showed recently (41) that the 362-peptide data set is simple to classify; an unsophisticated linear algorithm produces results (91.5% ± 3%) that no nonlinear algorithm beats with statistical significance, and the resulting classifier is simple to extract specificity rules from. It is therefore rather pointless to continue to apply complicated (nonlinear) algorithms on the 362-peptide data set; the resulting models will be harder to interpret than the linear models but no better in classification performance. We have therefore collected a new data set with 746 peptides, more than twice the size of the old data set, and used a number of approaches to predict cleaving and extract rules. This has allowed us to draw some new conclusions about HIV-1 protease cleavage specificity but also to discover that the problem remains linear when represented in the standard way.

(Preliminary results from this study were reported at the ISMB/ECCB Conference in Glasgow, Scotland, in 2004 [53].)

MATERIALS AND METHODS

Collecting the sequence data. The literature was searched for publications containing experimental data on oligopeptide sequences that have been exposed to the HIV-1 protease (3–5, 24, 25, 27, 29, 32, 38, 39, 45–48). References used previously by Cai and Chou (11) were excluded since their data set was added at a later stage (see below). Oligopeptides from the references were labeled as "cleaved," "uncleaved," or "undetermined," depending on the information given in the reference. "Undetermined" oligopeptides were those where the k_{cat}/K_m value had not been determined (K_m is the Michaelis constant, the substrate concentration at which the reaction rate is half the maximal; k_{cat} is the turnover number, the number of substrate molecules converted into product per unit of time at a single catalytic site when the enzyme is fully saturated with substrate; and k_{cat}/K_m is the rate constant for the interaction of substrate and enzyme). "Uncleaved" oligopeptides were those that had been labeled as "not hydrolyzed" or as "uncleaved." "Cleaved" oligopeptides were those that had been labeled as "cleaved." Three oligopeptides were listed by Beck et al. (4) in two slightly different ways, as "<5% cleavage" and as "not cleaved," which we labeled as uncleaved. These were KGSGVY*QLSALVPK, KGSGGRIN*VALVPK and KSGVF*SVNGLVK, where the scissile bonds are marked with asterisks. One conflict occurred among the collected oligopeptides: the oligopeptide TERQA N*FLGKI was labeled as "not hydrolyzed" by Tözsér et al. (46) and "cleaved" by Kurt et al. (29). Since RQAN*FLGK is the known NC/p1 cleavage site, we designated this oligopeptide as "cleaved." "Undetermined" oligopeptides were removed from the data set.

It is believed that only eight residues (P_4 to $P_{4'}$) are important for the HIV-1 protease specificity, which is why octamers were extracted from the cleaved and uncleaved oligopeptides. Cleaved octamers were extracted such that the cleavage site was located between P_1 and $P_{1'}$; e.g., the cleaved oligopeptide SGLTM*V QELV (3) resulted in the octamer GLTMVQEL. Uncleaved octamers were extracted from uncleaved oligopeptides by using an eight-residue-long sliding window. For example, the uncleaved oligopeptide KSGVFVNNGLVK (4) generates the five uncleaved octamers KSGVFVNN, SGVFVNNG, GVFVNNGL, VFVNNGLV, and FVNNGLVK. This resulted in 412 octamers (308 cleaved and 104 uncleaved).

The 412 octamers were then joined with the 362 octamers listed previously by Cai and Chou (11), which resulted in a few conflicts; i.e., the same octamer was labeled as both cleaved and uncleaved in different publications. Such conflicting octamers were removed from the data set, with one exception: the octamer RQNYPIVQ was labeled as "uncleaved" by Cai and Chou (11) and as "cleaved" by Tözsér et al. (45). This octamer was kept and labeled as "cleaved" since it is so similar to the known MA/CA cleavage site SQNYPIVQ in the Gag-Pol polyprotein.

This procedure, summarized in Fig. 1, yielded a final data set with 746 octamers (401 cleaved and 345 uncleaved).

Sequence representation. The octamers are eight-character-long "words" that need to be represented numerically to the algorithm. How this is done can be of crucial importance for the result. We used three different representations, depending on the model used: similarity matrix distance, sparse orthogonal coding scheme, and property coding. The similarity matrix representation, where the octamer is compared to other octamers using similarity matrices, was used for the biobasis networks (44). The property coding was used together with nonlinear support vector machines only. The sparse orthogonal coding scheme (36) was used for all other algorithms.

Each amino acid is in the sparse orthogonal coding represented by a 20-bit binary vector, where 19 bits are set to zero and 1 bit is set to 1, in such a way that the scalar product of two different amino acid vectors is zero. This coding is standard when neural networks are used for sequence analysis (2). The property coding, on the other hand, builds on replacing each amino acid with one or more properties, e.g., "hydrophobic," "small," "polar," and so on. We used the properties listed in Table 1, where the "hydrophobic" property is based on scaled hydrophobicity values (7) taken from the website http://psyche.uthct.edu/shaun/SBlack/aagrease.html, and the others are based on the Venn diagram on the website http://www.russell.embl-heidelberg.de/aas/aas.html (6). Property coding transforms the octamer to a binary string with 8, 16, or 24 bits depending on whether one, two, or three properties are used, respectively. For example, the

TABLE 1. Properties for amino acids

Property	Amino acids
Small	G, C, N, S, T, D, A, V, P
Hydrophobic	A, V, P, M, F, L, Y, I, W, C
Charged	D, E, H, K, R
Polar	T, C, S, N, D, Q, E, K, R, H, Y, W
Aromatic	F, Y, W, H
Aliphatic	V, I, L

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

octamer SQNYPIVQ is recoded to 00011110 if only "hydrophobic" is used and to 0001111010101010 if both "hydrophobic" and "small" are used.

Bioinformatic algorithms. The description of bioinformatic algorithms is kept brief since details on algorithms can be found in standard textbooks (see, e.g., references 2, 18, and 25). We make a distinction between linear and nonlinear classification algorithms. Linear algorithms classify the octamer using a monotonic function $f(w \cdot x)$, where x is a representation of the octamer (see above), w is a set of parameters for the classifier, and the dot denotes a linear combination of w and x, e.g., a scalar product. Nonlinear algorithms classify the octamer with nonmonotonic functions f and/or with functions that take a nonlinear combination of w and x as argument. Linear algorithms are easier to interpret and in general much faster to train (calibrate) than nonlinear algorithms. We have used two linear algorithms and three nonlinear algorithms in this study. The linear ones were the simple perceptron (SP) and the linear support vector machine (SVM). The nonlinear ones were neural networks, support vector machines with Gaussian kernel, and biobasis networks.

The simple perceptron (26, 42) is a linear method that uses the perceptron training algorithm, which minimizes the number of mistakes made by the algorithm. A particular strength of this algorithm is that it is guaranteed to find a linear solution if one exists. The neural network algorithm is an extension of the simple perceptron to the so-called multilayer perceptron (MLP) (2, 26). The MLP is trained using a gradient descent type learning algorithm that minimizes the difference between the algorithm's predictions and the true answer (this is different from minimizing the number of mistakes). The MLP contains internal units, denoted hidden units, which control how nonlinear the algorithm can be. The nonlinearity of the MLP increases with the number of hidden units, and an MLP with one hidden unit is a linear algorithm.

The SVM uses a kernel function to map the octamer x into a high-dimensional kernel space where the problem is easier to solve (18). The simplest SVM uses a linear kernel, i.e., it is just a linear classifier, whereas more complicated SVMs can use Gaussian or polynomial kernels. We used a linear kernel together with the sparse orthogonal coding scheme and a Gaussian kernel for the property-coding scheme.

The biobasis network (44) is a recently suggested algorithm based on protein similarity matrices where an octamer, x, is classified according to its evolutionary distance from type patterns in a database. The type patterns are expected to capture important characteristics of the data set. A problematic issue with the biobasis network is which set of similarity matrices to use; PAM250 is a popular one in the literature, but different matrices will yield different results. Since our data set only contains octamers and HIV-1 protease only recognizes specific cleavage sites, other percent accepted mutation (PAM) matrices with short evolutionary distance might be more appropriate for this data set. We therefore tried different PAM matrices for our biobasis experiments.

All the algorithms were run using the MATLAB (http://www.mathworks.com) numerical computing tool and using the Mathworks neural network toolbox for the simple perceptron and the MLPs, our own implementation for the biobasis functions, and the OSU SVM 3.0 toolbox for the SVMs (http://www.ece.osu.edu/~maj/osu_svm/).

Validation of the methods. It is important to separate in-sample and out-of-sample performance. The in-sample performance is how well an algorithm can learn data; it is often the case that a good algorithm can learn all data perfectly. The in-sample performance is, however, usually not very interesting; the out-of-sample performance is the important result. The out-of-sample performance is how well the algorithm can generalize its knowledge to data that it has not previously seen. The out-of-sample performance for all methods was measured using the following cross-validation scheme: the data set was randomly divided into two parts, of which one part was used to construct the classifier and the other part was used to test the classifier. This was repeated 100 times, and the final reported performance was the average of all 100 tests. This process also produced variance estimates for the performance so that the significance of differences between algorithms could be measured.

RESULTS

Structure of the data set with sparse orthogonal coding. The first experiment was to use a simple perceptron to test if the 746-octamer data set could be linearly separated when the sparse orthogonal encoding is used, since the smaller data set with 362 octamers is known to be linearly separable in this encoding. The data set was found to be linearly separable, which is interesting when compared to the probability for it to be a random event. The capacity of a linear classifier (26) yields that the probability (the P value) for 746 random peptides to be linearly separable is less than 10^{-58}, i.e., extremely low. It is thus highly surprising to find that our data set is linearly separable, and it must say something about either how data were originally collected (in the virus laboratory) or the biochemistry of the problem. It also hints that it is unlikely that a nonlinear algorithm will be better than a linear algorithm when the sparse orthogonal encoding is used with this data set. The structure and linear separation of the data set with the sparse orthogonal encoding are shown in Fig. 2.

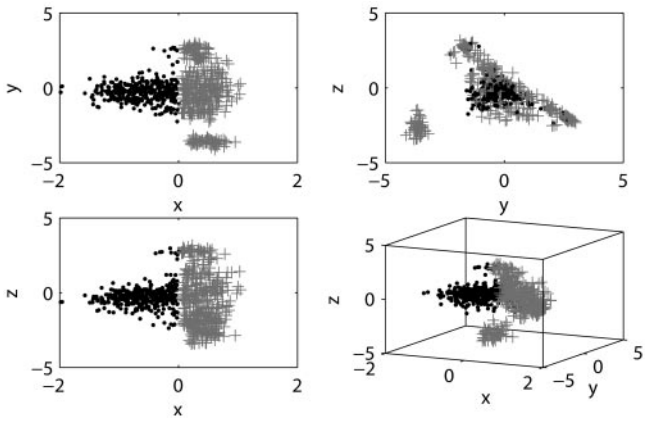

FIG. 2. Plot showing the structure of the 746-peptide data set with sparse orthogonal encoding (uncleaved peptides are black dots, and cleaved peptides are gray crosses). The peptides have been mapped in three directions: the direction along which the data set is linearly separable (x), the direction orthogonal to x along which the data set has the largest variation (y), and the direction orthogonal to x and y along which the data set has the second largest variation (z).

Structure of the data set with property coding. The second experiment was done to check how well cleaved and uncleaved octamers were separated by different amino acid properties. This was done by combining one or two properties from Table 1, recoding the octamers according to these properties, and then measuring the overlap between the cleaved and the uncleaved categories. The overlap was defined as the number of unique patterns that, after property coding, occurred within both the cleaved and the uncleaved categories. The results, summarized in Table 2, showed that the size and hydrophobicity properties in combination separated the data very well and that these two properties might even be sufficient to describe HIV-1 protease specificity.

The significance of the best results in Table 2 was checked with a computer simulation to estimate the risk that the result was just a random event. There are 9 "small" and 10 "hydrophobic" amino acids. We tested the significance of the low overlap (1.7%) by randomly labeling 9 amino acids as "small" and 10 amino acids as "hydrophobic" and computing the resulting overlap after this random assignment of two properties. The random labeling was repeated 1 million times, after which we counted how often the overlap was less than or equal to 1.7%. This happened 35,957 times, and the low overlap displayed by the properties "small" and "hydrophobic" is therefore significant at a level of 3.6%. The corresponding number for the second best property pair, "small" and "polar," was

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

TABLE 2. Cleaved/uncleaved overlap when different properties are used

Property	No. of unique cleaved sites	No. of unique uncleaved sites	No. of overlap sites (%)
Small	95	160	63 (32.8)
Hydrophobic	91	170	59 (29.2)
Charged	46	81	34 (36.6)
Polar	112	171	76 (36.7)
Aromatic	26	39	19 (41.3)
Aliphatic	60	78	42 (43.8)
Small and hydrophobic	245	338	10 (1.7)
Small and polar	254	340	12 (2.1)
Small and aliphatic	236	313	19 (3.6)
Hydrophobic and aliphatic	190	282	26 (5.8)
Charged and polar	189	302	29 (6.3)

97,342 times. The "polar" property is, however, almost complementary to "hydrophobic," and we therefore considered only the better of these two pair combinations in the later analysis.

We also tested if the property-coded data set could be linearly separated and found that this was not the case; the HIV-1 protease cleavage specificity problem is nonlinear when property coding is used.

Cleavage prediction. The third experiment was to build predictors that predicted whether an octamer would be cleaved or not by the HIV-1 protease. All the different algorithms mentioned in Materials and Methods were tested in this experiment, i.e., SP, MLP, linear SVM, Gaussian SVM (GSVM), and BBF with different similarity matrices. The out-of-sample prediction performance of all algorithms was tested with cross-validation. In a previous paper (41), we concluded that the prediction performance for MLPs decreases if the number of hidden nodes increases. We therefore only considered MLPs with two hidden units (the smallest number that still yields a nonlinear algorithm) here.

Figure 3 shows the out-of-sample prediction performance of the methods that turned out to work the best plotted versus the size of the data set used to construct the predictor (referred to as the training data). The error bars are the corresponding one-sided 95% confidence interval for the values. The GSVM with property coding worked better than the other methods with sparse orthogonal coding, although the difference was not significant (at the 95% level) between the GSVM model and the best model using orthogonal coding. The BBFs performed very poorly (not shown) and were excluded from the figure and further analysis. Care was taken in this experiment to make sure that no two sequences with the same property coding were mixed between the training data and the out-of-sample test data, since this would have biased the result towards better classification performance (this bias effect is illustrated in the figure).

The predictors were also tested on sequence data that were completely out of sample, i.e., data not included in the 746-peptide set. The website http://www.bioafrica.net (22) lists several HIV-1 cleavage sites, of which about half are not in our data set. Table 3 lists some of these cleavage sites together with the SP with sparse orthogonal encoding and GSVM with property-coding classifier predictions. The simple perceptron correctly classified seven out of eight previously unseen cleavage sites, and the GSVM correctly predicted all eight to be cleaved. Furthermore, the sequence that was erroneously classified by the simple perceptron as uncleaved was very close to being classified as cleaved.

de Oliveira et al. (22) also listed several variations on cleavage sites in their paper. It was possible to generate 160 octamers from those lists, of which 140 were not in the 746-peptide set. The predictors were therefore tested on these 140 octam-

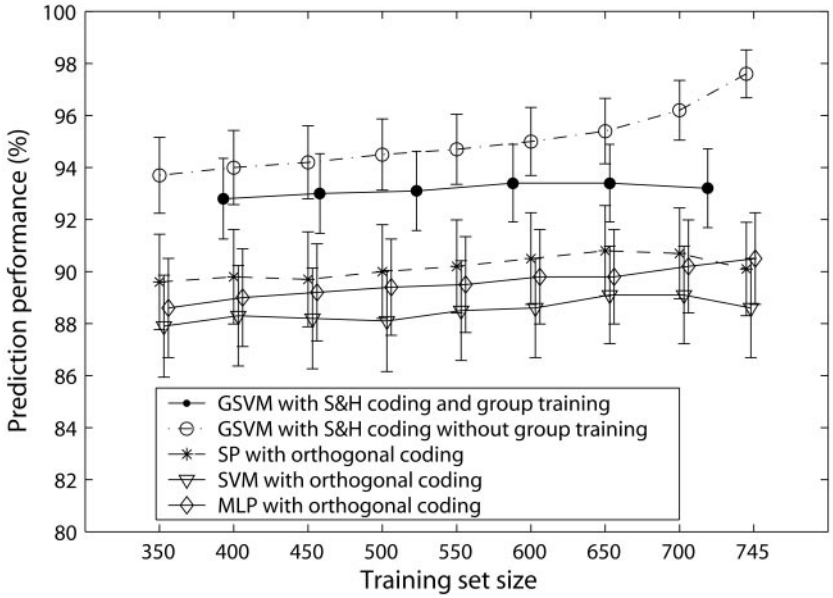

FIG. 3. The best predictors' out-of-sample performances, estimated using cross-validation. The two upper curves are both for property-coded data, but the top one represents the case when care is not taken to avoid sequence bias in the testing (shown here to illustrate the importance of avoiding such bias and overly optimistic results). Here, S denotes small property and H denotes hydrophobicity.

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

TABLE 3. Some known HIV-1 protease cleavage sites and SP and GSVM predictions for these sites

Sequence	Site	In 746-peptide set?	SP prediction	GSVM prediction
SQNYPIVQ	MA/CA	Yes	C	C
ARVLAEAM	CA/p2	Yes	C	C
TAIMMQKG	p2/NC	No	C	C
SAIMMQRG	p2/NC	No	C	C
RQANFLGK	NC/p1	Yes	C	C
PGNFLQSR	p1/p6^{gag}	Yes	C	C
RQANFLRE	NC/TFP	No	C	C
NLAFQQGE	TFP/p6^{Pol}	No	C	C
DLAFLQKG	TFP/p6^{Pol}	No	C	C
SFSFPQIT	p6^{pol}/PR	No	C	C
SFNFPQVT	p6^{pol}/PR	Yes	C	C
TLNFPISP	PR/RT^{p51}	Yes	C	C
AETFYVDG	PR/RT^{p66}	Yes	C	C
RKVLFLDG	RT^{p66}/INT	Yes	C	C
DCAWLEAQ	Nef	No	C	C
ACAWLEAQ	Nef	No	NC	C

ers and the results, shown in Table 4, were similar to the cross-validation results (Fig. 3); the GSVM predictor performs the best, and the nonlinear classifiers are not better than the linear classifiers when sparse orthogonal coding is used.

A weakness with the results shown in Table 4 is that they include only cleaved sequences; a predictor that predicts all sequences to be cleaved would get 100% prediction accuracy on this test (but a *P* value of 1). A further test was therefore made using 78 octamers described previously by Pettit et al. (33), of which 62 (18 cleaved and 44 uncleaved) were not in the 746-peptide training data set. However, a problem with these data turned out to be that a slightly different definition was used than what we had used for cleaved/uncleaved octamers, since the 746-peptide data set contains four cleaved octamers that were defined as uncleaved by Pettit et al. (33): ARVAA EAM, ARVIAEAM, ARVNAEAM, and PGNLLQSR. The first three were collected previously (38), and the fourth one was reported previously by Fehér et al. (24). All four had low k_{cat}/K_m values in these papers; the experimental data were therefore not in conflict with the data reported previously by Pettit et al., but we defined octamers with low but nonzero k_{cat}/K_m values as cleaved, whereas Pettit et al. defined them to be uncleaved if they were not cleaved at a rate of at least 3 to 5% of that of the wild-type reference. A consequence of this difference in definition was that the models performed poorly, i.e., not significantly better than random, on uncleaved octamers described previously (33) and predicted too many to be cleaved. The performance on cleaved octamers was similar to that of the results shown in Table 4.

TABLE 4. Different predictors' performance on 140 cleaved peptides

Model	% Prior probability for cleaving[a]	% Prediction accuracy (no. of peptides/total no.)	Rank[b]	*p* value[c]
GSVM	17.8	98 (137/140)	1	10^{-98}
LSVM	15.0	94 (132/140)	2	10^{-97}
MLP com[d]	11.1	94 (131/140)	3	10^{-112}
SP	10.7	92 (129/140)	3	10^{-110}
MLP	13.1	91 (128/140)	3	10^{-97}

[a] How often a random sequence is predicted to be cleaved by the predictor.
[b] Predictors that are not significantly different are given the same rank, tested using McNemar's test (40) with a significance level of 95%.
[c] Probability for the observed prediction accuracy, or better, if the cleaving is predicted randomly.
[d] Committee of 100 MLP models, i.e., 100 MLP models that form a combined prediction.

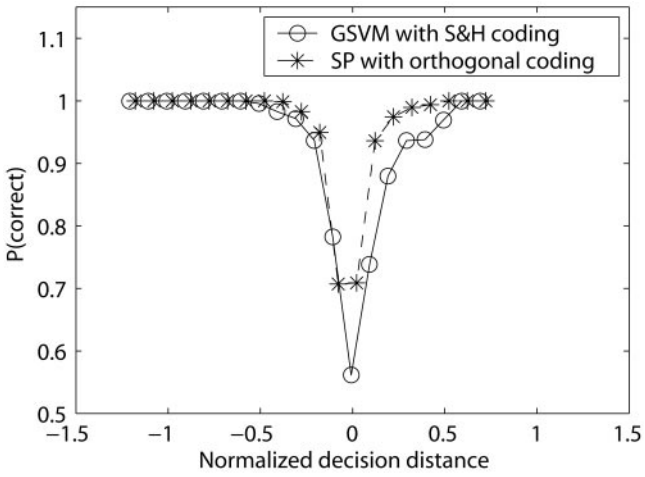

FIG. 4. The probability for the GSVM and SP predictors to be correct as a function of the distance from the boundary separating cleaved and noncleaved patterns. Negative distance denotes distance on the noncleaved side of the boundary, and positive distance denotes distance on the cleaved side of the boundary. The distance measures produced by the SP predictor and the GSVM predictor have been normalized so that they can be compared to each other. "S&H" denotes that property coding based on the properties "small" and "hydrophobic" was used.

In addition to a cleavage/noncleavage decision, a predictor also produces a measure of how close it is to changing its decision, and one must put less confidence in the prediction if the observed octamer lies close to the decision boundary, the boundary separating cleaved and noncleaved patterns. Figure 4 shows how the probability for the predictors to be correct (estimated using cross-validation) varies with the distance from the decision boundary. This relationship can be used to produce a confidence estimate for the prediction. If an octamer lies close to the boundary, the probability that the predictor is correct is just about 50%, but the probability for being correct increases quickly with the distance from the boundary.

It was found in a previous study of this problem (41) that an SP predictor was more correct when predicting cleavage for oligopeptides with high k_{cat}/K_m values than when predicting cleavage with low values. This result was, however, achieved with a very small out-of-sample test; we therefore tested if this was the case for the SP and GSVM predictors when cross-validation was used on our larger data set. The result, shown in Fig. 5, was that it is not possible to conclude that a predictor will be more correct for high k_{cat}/K_m values than for low k_{cat}/K_m values.

Screening potential HIV-1 protease cleavage sites. The present in silico model may perhaps prove useful to screen potential protein substrates, such as Gag and the Gag-Pol polyprotein, for cleavage sites for the HIV-1 protease. However, since the predictors in this study are solely based on how well HIV-1 protease cleaves short peptides and do not take

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

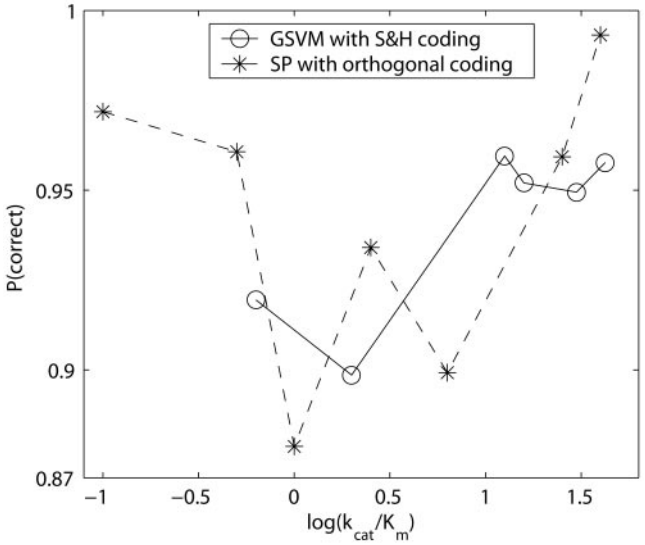

FIG. 5. Plot showing the probability for the predictors to be correct in their prediction versus the logarithm of the k_{cat}/K_m value for the oligopeptide. The error bars (95% confidence intervals) for the point estimates are approximately 10% (left out of the plot for readability reasons). The k_{cat}/K_m values for the point estimates were chosen such that roughly the same number of observations was used for each estimate. "S&H" denotes that property coding based on the properties "small" and "hydrophobic" was used.

into account the complexity of complete proteins, the results of such searches must be interpreted with great caution, because in vivo, the three-dimensional structure of a protein and post-translational modifications, such as phosphorylation and glycosylation, may interfere and hinder efficient cleavage. The in silico model will therefore predict more cleavage sites than are actually experimentally observed, but it should not fail to predict experimentally observed cleavage sites. This is also what we observed for Gag and Gag-Pol polyproteins.

We provide two further illustrations to this: human NDR1 serine/threonine kinase and eukaryotic initiation factor 4 gamma 1 (eIF4GI). An HIV-1 protease cleavage site was recently identified in NDR1 (23). The sequence is KDWVF/IN YTYKRFEGLTARGAIPSYMKAAK, with "/" denoting the experimentally identified cleavage site. The GSVM model predicts four potential cleavage sites in NDR1 (Table 5), but the correct cleavage site is much more certain than the others. The GSVM predictor thus supports the experimentally determined cleavage site for NDR1.

Three HIV-1 protease cleavage sites have been reported for human eIF4GI (49). These are shown in Table 5 together with the GSVM predictor's opinion for these sites. The GSVM model predicts several potential cleavage sites in eIF4GI but agrees with only one of the experimentally determined sites and, above all, disagrees strongly with one. The cleavage site that the GSVM disagrees strongly with is KIIA/TVLM, which has a small hydrophobic residue at site P_1 and a small nonhydrophobic residue at site $P_{1'}$. This is different from any previously observed cleavage sites (cf. position property preferences in Table 6), which makes it an important site to study further.

Extracting rules from the best predictor. The best predictor, the nonlinear Gaussian SVM, was used to generate specificity rules in the following way: there are 2^{16} (65,536) possible property patterns (the "size and hydrophobicity" representation contains 16 bits), which can be processed with the Gaussian SVM predictor in a few seconds. Individual bits, or combinations of bits, were therefore fixed, and the cleavage probability with fixed bits was compared to that without fixed bits. If the conditional probability for cleaving when the bits were fixed was much higher than the probability for cleaving with no fixed bits, that particular set of bits (properties) was deemed important for the cleaving. The results in Tables 6, 7, and 8, were generated with this method. The property preferences for the eight different sites are listed in Table 6, the 10 most important single position cleavage/noncleavage rules are listed in Table 7, and the 20 most important cleavage/noncleavage rules for two positions are listed in Table 8 (similar tables can be generated for more than two positions). The following notation is used for each position: S denotes small, H denotes hydrophobic, and ^ denotes "not." For example, P_1 = (^SH) means that position P_1 is fixed with a nonsmall and hydrophobic residue. The p_r value is the ratio between the conditional probability for cleaving, i.e., with the particular motif fixed, and the prior probability for cleaving, i.e., when nothing is fixed. For example, p_r = 3.0 for P_1 = (^SH) means that it is three times more likely that a sequence with this motif is predicted to be cleaved than a random sequence. The p_c value is the conditional probability for cleaving if the octamer contains the specific motif; e.g., p_c = 44.5% for P_1 = (^SH) means that 44.5% of all octamers that contain a nonsmall hydrophobic residue in position P_1 are predicted to be cleaved.

As seen in Table 6, position P_1 is the most specific: the probability for cleavage is increased threefold if the site is occupied by a nonsmall and hydrophobic residue; all other property combinations in P_1 decrease the probability for cleaving by more than half. The preferred residues in P_2 should be hydrophobic and nonsmall, although size is less important than hydrophobicity. The preferred residues in P_3 should be nonsmall. The preferred residues in P_4 should, on the other hand, be small. Hydrophobicity is unimportant for both P_4 and P_3. The $P_{1'}$ position should have a hydrophobic residue, large or small, and the $P_{2'}$ position should be occupied by anything but a small and nonhydrophobic residue. The $P_{3'}$ position is similar to position P_3.

Similar rule tables can also be generated from the SP pre-

TABLE 5. GSVM prediction of HIV-1 protease cleavage of NDR1 and eIFG1

NDR1 sequence with predicted cleavage sites	KDWVF / IN / YTY/KRFEGLTARGAIPS/YMKAAK		
GSVM P(correct)[a]	0.92 0.58 0.59	0.64	
eIF4G1 sites	. . .KIIA/TVLM. . .	. . .ATVL/MTED. . .	. . .RFSA/LQQA. . .
GSVM prediction	Not cleaved	Cleaved	Not cleaved
GSVM P(correct)	0.95	0.70	0.63

[a] P(correct) is the estimated probability for the predictor to be correct (cf. Fig. 4).

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

TABLE 6. Property preferences for the eight substrate positions

P_4		P_3		P_2		P_1		$P_{1'}$		$P_{2'}$		$P_{3'}$		$P_{4'}$	
Motif	p_r[a]	Motif	p_r	Motif	p_r	Motif	p_r	Motif	p_r	Motif	p_r	Motif	p_r	Motif	p_r
^S^H	0.9	^S^H	1.4	^S^H	0.4	^S^H	0.4	^S^H	0.4	^S^H	1.4	^S^H	1.1	^S^H	1.0
^SH	0.6	^SH	1.3	^SH	1.2	^SH	3.0	^SH	1.9	^SH	1.3	^SH	1.4	^SH	0.9
S^H	1.3	S^H	0.8	S^H	0.7	S^H	0.2	S^H	0.3	S^H	0.4	S^H	0.8	S^H	1.3
SH	1.2	SH	0.6	SH	1.6	SH	0.3	SH	1.3	SH	1.0	SH	0.7	SH	0.8

[a] Conditional probability for the sequence to be cleaved divided by the prior probability.

dictor using the orthogonal representation (41), and the most prominent motifs are shown in Table 9 for reference.

Interdependencies among the P_i positions. Ridky et al. (38) reported on observed interdependencies between positions in the substrate and showed that there is a dependency between, e.g., the P_1 and P_3 positions; large residues in both these positions lead to less efficient cleavage than when there is a large residue in only one of the positions. It is difficult to translate this to the GSVM predictor since it does not reflect k_{cat}/K_m values (Fig. 5). However, the nonlinearity of the HIV-1 protease problem, when size and hydrophobicity are used, indicates that there is interaction between positions. This is because when using a binary notation, nonlinearity means having a function of the form "exclusive or" (XOR), which is a Boolean logic function that takes the value as true when either of its two inputs is true but not both. An example of an XOR relationship is if we find that a large residue in position P_1 or a large residue in position P_3 leads to cleavage but that large residues or small residues in both positions lead to no cleavage. Thus, the presence of an XOR relationship means that there is competitive interaction between positions.

The GSVM model was therefore probed for such XOR relations. This was done by selecting two positions, fixing the remaining 14 positions, and computing the cleaving prediction when the two selected positions took the values 00, 01, 10, and 11. If the corresponding cleaving prediction was "not cleaved," "cleaved," "cleaved," and "not cleaved" (or the inverse of this), respectively, then there was an XOR relationship between the selected pair of positions. This was repeated for all possible 2^{14} (16,384) values for the remaining 14 positions, counting the relative number of times an XOR relation was encountered, and it was taken as a sign of a strong competitive relationship between the two chosen positions if XOR relations occurred frequently. This was done for all positions and for both size and hydrophobicity properties.

We found that occasional XOR relationships occurred for most pairs of positions regarding residue size; i.e., it was almost always possible to find a certain amino acid sequence for which a pair of amino acids would exhibit XOR behavior with respect to size (or hydrophobicity). When all possible sequences were tested, it was found that pairs that exhibited size-related XOR interaction more often than others (albeit less than 1% of the time) were $P_{3'}$ and $P_{4'}$, P_1 and $P_{4'}$, and $P_{1'}$ and $P_{2'}$. Pairs that exhibited hydrophobicity XOR interaction more often than others (again, less than 1% of the time) were $P_{3'}$ and $P_{4'}$, P_4 and $P_{3'}$, P_4 and P_3, and $P_{2'}$ and $P_{3'}$. The specific presence of *cis* or *trans* position interactions (38) could thus not be inferred from the structure of the GSVM model.

DISCUSSION

We have reapproached the HIV-1 protease specificity problem by compiling a new, extensive collection of published octamers that are cleaved or not cleaved by the HIV-1 protease and have analyzed these data with several popular bioinformatic machine learning algorithms. The new data represent a significant amount of know-how about the HIV-1 protease: it is a summary of published experimental results from different laboratories over the last 15 years. Our aim with this work has been both to build the best predictors for HIV-1 protease specificity and to analyze HIV-1 cleaving characteristics.

It was surprising to find the data set to be linearly separable when sparse orthogonal encoding was used (Fig. 2). The probability that this would happen by chance is so extremely low that it must say something about how the data were originally collected and/or the biochemistry of the problem. A problem associated with the sampling of cleaved sequences is that they are often collected by slightly varying the amino acid sequence of a known cleavage site and exposing the resulting oligopeptide to HIV-1 protease. The sampling of the cleaved octamer space is thus not especially random, and the variation in the data set is smaller than what the number of observations indicates. On the other hand, this also leads to many of the uncleaved octamers having a high sequence similarity with cleaved octamers, which does not make the problem any simpler. Furthermore, the variation between the HIV-1 cleavage

TABLE 7. The 10 most important single-position motifs for cleaving and noncleaving

Motif with high probability for cleavage	p_c (%)[a]	Motif with high probability for cleavage	p_c (%)	Motif with low probability for cleavage	p_c (%)	Motif with low probability for cleavage	p_c (%)
P_1 = (^SH)	44.5	$P_{3'}$ = (^SH)	20.1	P_1 = (S^H)	3.6	P_2 = (^S^H)	6.5
$P_{1'}$ = (^SH)	28.3	$P_{4'}$ = (S^H)	19.5	P_1 = (SH)	4.2	P_1 = (^S^H)	6.5
P_2 = (SH)	23.4	$P_{1'}$ = (SH)	19.2	$P_{1'}$ = (S^H)	5.1	P_3 = (SH)	8.1
P_3 = (^S^H)	20.8	P_4 = (S^H)	19.0	$P_{2'}$ = (S^H)	5.7	P_4 = (^SH)	9.0
$P_{2'}$ = (^S^H)	20.3	P_3 = (^SH)	18.6	$P_{1'}$ = (^S^H)	6.3	P_2 = (S^H)	10.7

[a] Conditional probability for the sequence to be cleaved (prior probability for any sequence is p_c = 14.7%).

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

TABLE 8. The 20 most important pair combination motifs for cleaving and noncleaving

Motif with high probability for cleavage	p_c (%)	Motif with high probability for cleavage	p_c (%)	Motif with low probability for cleavage	p_c (%)	Motif with low probability for cleavage	p_c (%)
P_1 = (^SH)	74.8	P_1 = (^SH)	53.1	P_2 = (^S^H)	0.2	P_2 = (S^H)	0.6
$P_{1'}$ = (^SH)		$P_{3'}$ = (^SH)		P_1 = (SH)		P_1 = (S^H)	
P_1 = (^SH)	61.4	P_1 = (^SH)	48.4	P_1 = (S^H)	0.3	P_1 = (SH)	0.7
$P_{1'}$ = (SH)		$P_{3'}$ = (^S^H)		$P_{1'}$ = (S^H)		$P_{2'}$ = (S^H)	
P_2 = (SH)	60.4	P_4 = (SH)	48.3	P_2 = (^S^H)	0.4	P_1 = (^S^H)	0.8
P_1 = (^SH)		P_1 = (^SH)		P_1 = (^S^H)		$P_{2'}$ = (S^H)	
P_1 = (^SH)	60.0	P_1 = (^SH)	46.4	P_1 = (^S^H)	0.4	P_1 = (SH)	0.8
$P_{2'}$ = (^S^H)		$P_{4'}$ = (^S^H)		$P_{1'}$ = (^S^H)		$P_{1'}$ = (S^H)	
P_4 = (S^H)	57.7	P_2 = (SH)	43.1	P_1 = (^S^H)	0.4	P_1 = (S^H)	0.8
P_1 = (^SH)		$P_{1'}$ = (^SH)		$P_{1'}$ = (S^H)		$P_{3'}$ = (SH)	
P_3 = (^SH)	57.3	P_3 = (^S^H)	41.7	$P_{1'}$ = (S^H)	0.4	P_1 = (SH)	0.9
P_1 = (^SH)		P_1 = (^SH)		$P_{2'}$ = (S^H)		$P_{1'}$ = (^S^H)	
P_3 = (^S^H)	56.9	$P_{1'}$ = (^SH)	38.5	P_2 = (^S^H)	0.5	P_2 = (^S^H)	0.9
P_1 = (^SH)		$P_{3'}$ = (^SH)		P_1 = (S^H)		$P_{1'}$ = (S^H)	
P_1 = (^SH)	54.2	$P_{1'}$ = (^SH)	37.1	P_2 = (^S^H)	0.5	P_1 = (S^H)	1.0
$P_{2'}$ = (^SH)		$P_{4'}$ = (S^H)		$P_{1'}$ = (^S^H)		$P_{2'}$ = (S^H)	
P_1 = (^SH)	54.1	P_2 = (^SH)	35.8	P_2 = (S^H)	0.5	P_3 = (S^H)	1.2
$P_{4'}$ = (S^H)		$P_{1'}$ = (^SH)		P_1 = (SH)		P_1 = (SH)	
P_2 = (^SH)	53.4	P_3 = (^S^H)	35.8	P_1 = (S^H)	0.6	P_3 = (SH)	1.3
P_1 = (^SH)		P_2 = (SH)		$P_{1'}$ = (^S^H)		P_1 = (SH)	

sites is large, and some of the cleaved sequences were collected using a bacteriophage library (3), which corresponds to a random sampling of hexamer space. It is therefore impossible to dismiss the linearity of the problem as an artifact of the data sampling; the linearity must also reflect the biochemistry. A possible interpretation is that the specificity is based on just a few amino acid properties, since few properties mean a low number of possible motifs and hence a higher probability for linear separation when the orthogonal encoding is used. This interpretation is supported by the finding that the cleaved and uncleaved octamers can be separated almost perfectly when only two properties, "small" and "hydrophobic" (or "polar"), are used, indicating that two properties may be sufficient to determine whether a given octamer will be cleaved or not. This is a new finding that, however, agrees well with previous experimental and structural results.

An observation in favor of this is the fact that mutations in the HIV-1 protease cleavage sites have a propensity for preserving membership in the groups "small" and "hydrophobic" (or "polar") along with their complements shown in Table 1 (17, 19, 30, 54). For example, the NC/p1 cleavage site RQAN FLGK often mutates to RQVNFLGK (17, 19, 54) in patients treated with protease inhibitors, and both A (Ala) and V (Val) are small hydrophobic amino acids. Several experimental results also indicate that hydrophobicity and size are important for HIV-1 protease specificity. For instance, work on molecular structure models supports the idea that size determines HIV-1 protease specificity rather than a particular amino acid sequence (35). It has also been observed that the effect of an HIV-1 protease inhibitor varies significantly with the hydrophobicity and size of the inhibitory molecules (16).

A major reason why the specificity of HIV-1 protease is interesting is the need to develop efficient protease inhibitors for clinical use. Inhibitors bind at the active site of the protease and compete with natural substrates, and we can compare the rules generated from the Gaussian SVM with current clinically used protease inhibitors. The strongest positive rule in Table 8 states that the protease prefers nonsmall and hydrophobic amino acids on both sides of the scissile bond. Saquinavir, ritonavir, lopinavir, and atazanavir all have bulky and hydrophobic structures in P_1 and $P_{1'}$. Ritonavir also has Val (V) in the P_2 position, which is consistent with the preference P_2 = (SH). An inconsistency, however, exists for saquinavir, which has the small polar Asn (N) in the P_2 position, which also is in

TABLE 9. The most important amino acid positions (single and pair) from the SP predictor

Sequence with high probability for cleavage	p_c (%)[a]	Sequence with high probability for cleavage	p_c (%)	Sequence with low probability for cleavage	p_c (%)	Sequence with low probability for cleavage	p_c (%)
xxxYxxxx	39	xxxYxExx	80	xxxxKxxx	1	xxxKKxxx	≅0
xxxxxExx	38	xxxMxExx	74	xxxKxxxx	1	xxKxKxxx	0.02
xxxMxxxx	32	xxxFxExx	73	xxKxxxxx	2	xxxxKKxx	0.02
xxxFxxxx	31	xxxYYxxx	72	xxxxSxxx	2	xxQxKxxx	0.02
xxxxYxxx	28	xxxxYExx	69	xxQxxxxx	2	xxxKSxxx	0.02
xExxxxxx	28	xExYxxxx	69	xxxxxKxx	2	xxxVKxxx	0.03
xxxxxVxx	26	xxxYxVxx	68	xxxVxxxx	3	xxKKxxxx	0.03
xxxxFxxx	25	xxxYFxxx	67	xxxxxxxI	3	xxxTKxxx	0.03
xxVxxxxx	24	xExxxExx	67	xxxTxxxx	3	xxxKxKxx	0.04
xFxxxxxx	23	xxxxFExx	65	Vxxxxxxx	4	xxQKxxxx	0.04

[a] Conditional probability for the sequence to be cleaved (prior probability for any sequence xxxxxxx is p_c = 11%).

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

agreement with recent experiments showing Asn at P_2 to be a preferred amino acid for cleavage (1). The Gaussian SVM says that a small nonhydrophobic residue at P_2, i.e., $P_2 = (S \hat{} H)$, should be negative for cleaving. The Gaussian SVM rules instead indicate that efforts should be spent on designing inhibitors that have a large hydrophobic amino acid in P_1 and a small hydrophobic amino acid in $P_{1'}$, since this is the second most promoting motif.

Our intention is that the specificity rules and the predictors for HIV-1 protease cleavage can be used to guide further research and compare identified cleavage sites with knowledge from previously identified cleavage sites. For instance, information from our previous work on computational prediction of HIV-1 protease cleavage specificity using the smaller data set of 362 peptides (41) was recently used in the identification process of potential protease cleavage sites in the human NDR1 and NDR2 serine/threonine kinases (23). When the Val-Phe (V-F) residues (P_2 and P_1 sites of the putative proteolysis site) were changed to Lys-Lys (K-K) (amino acids that have a very low probability of cleavage by the HIV-1 protease), mutant NDR1 and NDR2 kinases, in contrast to the corresponding wild-type kinases, were not processed in HIV-1-expressing cells (23). This is an indication of the predictive value and validity of our in silico model for future in vitro studies on HIV-1 protease. Furthermore, it lends support to the notion that lysine, especially the Lys-Lys motif, may be useful to incorporate in effective peptide-derived protease inhibitors, since it obviously leads to noncleavage of a previously cleavable substrate. In line with this, a recent paper identifies several lysine derivatives as potent HIV protease inhibitors which, because of their high specificity and low cytotoxicity, may be candidates for clinical trials (9). Interestingly, lysine residues may also be important in blocking viral entry, since a "template-assembled synthetic peptide" in which a lysine-rich decapeptide or hexapeptide, used as a template to covalently anchor five molecules of RPR, KPR, or RPK tripeptides (selected based on the RP dipeptide motif of the third hypervariable region [V3] of the extracellular envelope glycoprotein), gave over 85% inhibition of HIV entry in the human lymphoid CEM cell line, probably by binding to a cell surface component (13).

It is very interesting that our best predictor disagrees strongly with one of the identified cleavage sites in eIF4GI, indicating that this site (KIIA/TVLM) should be investigated further since it must contain completely new information about HIV-1 specificity.

The two best predictors, the GSVM property-based predictor and the linear simple perceptron with sparse orthogonal coding, will be available on our website (http://www.hh.se/staff/bioinf). The 746-peptide data set will also be available from this website.

ACKNOWLEDGMENTS

We thank Per-Åke Jovall and the anonymous reviewers for valuable feedback to this work.

L.Y. is funded within the National Research School in Genomics and Bioinformatics hosted by Göteborg University, Sweden. D.G. is the recipient of a postdoctoral scholarship from the Swedish Society for Medical Research and Government ALF funding for young clinical researchers from Lund University Hospital and the Faculty of Medicine at Lund University, Sweden.

REFERENCES

1. **Bagossi, P., T. Sperka, A. Fehér, J. Kádas, G. Zahuczky, G. Miklóssy, P. Boross, and J. Tözsér.** 2005. Amino acid preferences for a critical substrate binding subsite of retroviral proteases in type 1 cleavage sites. J. Virol. **79:**4213–4218.
2. **Baldi, P., and S. Brunak.** 2001. Bioinformatics—the machine learning approach, 2nd ed. MIT Press, Cambridge, Mass.
3. **Beck, Z. Q., L. Hervio, P. E. Dawson, J. H. Elder, and E. L. Madison.** 2000. Identification of efficiently cleaved substrates for HIV-1 protease using a phage display library and use in inhibitor development. Virology 2**74:**391–401.
4. **Beck, Z. Q., Y. C. Lin, and J. H. Elder.** 2001. Molecular basis for the relative substrate specificity of human immunodeficiency virus type 1 and feline immunodeficiency virus proteases. J. Virol. **75:**9458–9469.
5. **Beck, Z. Q., G. M. Morris, and H. J. H. Elder.** 2002. Defining HIV-1 protease substrate selectivity. Curr. Drug Targets Infect. Disord. **2:**37–50.
6. **Betts, M. J., and R. B. Russell.** 2003. Amino acid properties and consequences of substitutions, p. 289–316. *In* M. R. Barnes and I. C. Gray (ed.), Bioinformatics for geneticists, vol. I. Wiley, Chichester, United Kingdom.
7. **Black, S. D., and D. R. Mould.** 1991. Development of hydrophobicity parameters to analyze proteins which bear post- or cotranslational modifications. Anal. Biochem. **193:**72–82.
8. **Boden, D., and M. Markowitz.** 1998. Resistance to human immunodeficiency virus type 1 protease inhibitors. Antimicrob. Agents Chemother. **42:**2775–2783.
9. **Bouzide, A., G. Sauvé, and J. Yelle.** 2005. Lysine derivatives as potent HIV protease inhibitors. Discovery, synthesis and structure-activity relationship studies. Bioorg. Med. Chem. Lett. **15:**1509–1513.
10. **Brik, A., and C.-H. Wong.** 2003. HIV-1 protease: mechanism and drug discovery. Org. Biomol. Chem. **1:**5–14.
11. **Cai, Y. D., and K. C. Chou.** 1998. Artificial neural network model for predicting HIV protease cleavage sites in protein. Adv. Eng. Software **29**(2)**:** 119-128.
12. **Cai, Y. D., X. J. Liu, X. B. Xu, and K. C. Chou.** 2002. Support vector machines for predicting HIV protease cleavage sites in protein. J. Comput. Chem. **23:**267–274.
13. **Callebaut, C., E. Jacotot, G. Guichard, B. Krust, M. Rey-Cuille, D. Cointe, N. Benkirane, J. Blanco, S. Muller, J. Briand, and A. G. Hovanessian.** 1996. Inhibition of HIV infection by pseudopeptides blocking viral envelope glycoprotein-mediated membrane fusion and cell death. Virology **218:**181–192.
14. **Chen, L., A. Perlina, and C. J. Lee.** 2004. Positive selection detection in 40,000 human immunodeficiency virus (HIV) type 1 sequences automatically identifies drug resistance and positive fitness mutations in HIV protease and reverse transcriptase. J. Virol. **78:**3722–3732.
15. **Chou, K. C.** 1996. Prediction of human immunodeficiency virus protease cleavage sites in proteins. Anal. Biochem. **233:**1–14.
16. **Clemente, J. C., R. E. Moose, R. Hemrajani, L. R. Whitford, L. Govindasamy, R. Reutzel, R. McKenna M. Agbandje-McKenna, M. M. Goodenow, and B. M. Dunn.** 2004. Comparing the accumulation of active- and nonactive-site mutations in the HIV-1 protease. Biochemistry **43:**12141-12151.
17. **Côté, H. C. F., Z. Brumme, and P. R. Harrigan.** 2001. Human immunodeficiency virus type 1 protease cleavage site mutations associated with protease inhibitor cross-resistance selected by indinavir, ritonavir, and/or saquinavir. J. Virol. **75:**589–594.
18. **Cristianini, N., and J. Shawe-Taylor.** 2000. An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge, Mass.
19. **Dauber, D. S., R. Ziermann, N. Parkin, D. J. Maly, S. Mahrus, J. L. Harris, J. A. Ellman, C. Petropoulus, and C. S. Craik.** 2002. Altered substrate specificity of drug-resistant human immunodeficiency virus type 1 protease. J. Virol. **76:**1359–1368.
20. **De Clercq, E.**2004. Antiviral drugs in current clinical use. J. Clin. Virol. **30:**115–133.
21. **De Clercq, E.**2004. HIV-chemotherapy and -prophylaxis: new drugs, leads and approaches. Int. J. Biochem. Cell Biol. **36:**1800–1822.
22. **de Oliveira, T., S. Engelbrecht, E. J. van Rensburg, M. Gordon, K. Bishop, J. zur Megede, S. W. Barnett, and S. Cassol.** 2003. Variability at human immunodeficiency virus type 1 subtype C protease cleavage sites: an indication of viral fitness? J. Virol. **77:**9422–9430.
23. **Devroe, E., P. A. Silver, and A. Engelman.** 2005. HIV-1 incorporates and proteolytically processes human NDR1 and NDR2 serine-threonine kinases. Virology **331:**181–189.
24. **Fehér, A., I. T. Weber, P. Bagossi, P. Boross, B. Mahalingam, J. M. Louis, T. D. Copeland, I. Y. Torshin, R. W. Harrison, and J. Tözsér.** 2002. Effect of sequence polymorphism and drug resistance on two HIV-1 Gag processing sites. Eur. J. Biochem. **269:**4114–4120.
25. **Hazebrouck, S., V. Machtelinckx-Delmas, J. J. Kupiec, and P. Sonigo.** 2001. Local and spatial factors determining HIV-1 protease substrate recognition. Biochem. J. **358:**505–510.
26. **Hertz, J., A. Krogh, and R. G. Palmer.** 1991. Introduction to the theory of

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

neural computation, vol. 1. Lecture notes, Santa Fe Institute, Studies in the Sciences of Complexity. Addison-Wesley, Reading, Mass.
27. **Kadás, J., I. T. Weber, P. Bagossi, G. Miklóssy, P. Boross, S. Oroszlan, and J. Tözsér.** 2004. Narrow substrate specificity and sensitivity toward ligand-binding site mutations of human T-cell leukemia virus type 1 protease. J. Biol. Chem. **279:**27148–27157.
28. **Kaplan, A. H., J. A. Zack, M. Knigge, D. A. Paul, D. J. Kempf, D. W. Norbeck, and R. Swanstrom.** 1993. Partial inhibition of the human immunodeficiency virus type 1 protease results in aberrant virus assembly and the formation of noninfectious particles. J. Virol. **67:**4050–4055.
29. **Kurt, N., T. Haliloglu, and C. A. Schiffer.** 2003. Structure-based prediction of potential binding and nonbinding peptides to HIV-1 protease. Biophys. J. **85:**853–863.
30. **Maguire, M. F., R. Guinea, P. Griffin, S. Macmanus, R. C. Elston, J. Wolfram, N. Richards, M. H. Hanlon, D. J. T. Porter, T. Wrin, N. Parkin, M. Tisdale, E. Furfine, C. Petropolous, B.W. Snowden, and J.-P. Kleim.** 2002. Changes in human immunodeficiency virus type 1 Gag at positions L449 and P453 are linked to I50V protease mutants in vivo and cause reduction of sensitivity to amprenavir and improved viral fitness in vitro. J. Virol. **76:** 7398–7406.
31. **Narayanan, A., X. Wu, and Z. R. Yang.** 2002. Mining viral protease data to extract cleavage knowledge. Bioinformatics **18:**S5–S13.
32. **Pettit, S. C., J. Simsic, D. D. Loeb, L. Everitt, C. A. Hutchison III, and R. Swanstrom.** 1991. Analysis of retroviral protease cleavage sites reveals two types of cleavage sites and the structural requirements of the P1 amino acid. J. Biol. Chem. **266:**14539–14547.
33. **Pettit, S. C., G. J. Henderson, C. A. Schiffer, and R. Swanstrom.**2002. Replacement of the P1 amino acid of human immunodeficiency virus type 1 Gag processing sites can inhibit or enhance the rate of cleavage by the viral protease. J. Virol. **76:**10226–10233.
34. **Poorman, R. A., A. G. Tomasselli, R. L. Heinrikson, and F. J. Kézdy.**1991. A cumulative specificity model for protease from human immunodeficiency virus types 1 and 2, inferred from statistical analysis of an extended substrate data base. J. Biol. Chem. **22:**14554–14561.
35. **Prabu-Jeyabalan, M., E. Nalivaika, and C. A. Schiffer.** 2002. Substrate shape determines specificity of recognition for HIV-1 protease: analysis of crystal structures of six substrate complexes. Structure **10:**369–381.
36. **Qian, N., and T. J. Sejnowskij.**1988. Predicting the secondary structure of globular proteins using neural network models. J. Mol. Biol. **202:**865–884.
37. **Randolph, J. T., and D. A. De Goey.** 2004. Peptidomimetic inhibitors of HIV protease. Curr. Top. Med. Chem. **4:**1079–1095.
38. **Ridky, T. W., C. E. Cameron, J. Cameron, J. Leis, T. Copeland, A. Wlodawer, I. T. Weber, and R. W. Harrison.** 1996. Human immunodeficiency virus, type 1 protease substrate specificity is limited by interactions between substrate amino acids bound in adjacent enzyme subsites. J. Biol. Chem. **271:**4709–4717.
39. **Ridky, T. W., A. Kikonyogo, J. Leis, S. Gulnik, T. Copeland, J. Erickson, A. Wlodawer, I. Kurinov, R. W. Harrison, and I. T. Weber.** 1998. Drug-resistant HIV-1 proteases identify enzyme residues important for substrate selection and catalytic rate. Biochemistry **37:**13835–13845.
40. **Ripley, B.** 1996. Pattern recognition and neural networks. Cambridge University Press, Cambridge, United Kingdom.
41. **Rögnvaldsson, T., and L. You.** 2004. Why neural networks should not be used for HIV-1 protease cleavage site prediction. Bioinformatics **20:**1702–1709.
42. **Rosenblatt, F.** 1962. Principles of neurodynamics. Spartan Books, New York, N.Y.
43. **Thompson, T. B., K. C. Chou, and C. Zheng.** 1995. Neural network prediction of the HIV-1 protease cleavage sites. J. Theor. Biol. **177:**369–379.
44. **Thomson, R., T. C. Hodgman, Z. R. Yang, and A. K. Doyle.** 2003. Characterizing proteolytic cleavage site activity using bio-basis function neural networks. Bioinformatics **19:**1741–1747.
45. **Tözsér, J., A. Gustchina, I. T. Weber, I. Blaha, E. M. Wondrak, and S. Oroszlan.** 1991. Studies on the role of the S_4 substrate binding site of HIV proteinases. FEBS Lett. **279:**356–360.
46. **Tözsér, J., I. Bláha, T. D. Copeland, E. M. Wondrak, and S. Oroszlan.** 1991. Comparison of the HIV-1 and HIV-2 proteinases using oligopeptide substrates representing cleavage sites in Gag and Gag-Pol polyproteins. FEBS Lett. **281:**77–80.
47. **Tözsér, J., P. Bagossi, I. T. Weber, J. M. Louis, T. D. Copeland, and S. Oroszlan.** 1997. Studies on the symmetry and sequence context dependence of the HIV-1 proteinase specificity. J. Biol. Chem. **272:**16807–16814.
48. **Tözsér, J., G. Zahuczky, P. Bagossi, J. M. Louis, T. D. Copeland, S. Oroszlan, R. W. Harrison, and I. T. Weber.** 2000. Comparison of the substrate specificity of the human T-cell leukemia virus and human immunodeficiency virus proteinases. Eur. J. Biochem. **267:**6287–6295.
48a. **UNAIDS.** 2040. Global estimates of HIV and AIDS as of end 2003, p. 10. *In* 2004 report on the global AIDS epidemic, 4th global report. UNAIDS, Geneva, Switzerland. [Online.] http://www.unaids.org.
48b. **UNAIDS/WHO.** 2004. Global summary of the AIDS epidemic December 2004, p. 1. *In* AIDS epidemic update 2004. UNAIDS/WHO, Geneva, Switzerland. [Online.] http://www.unaids.org.
49. **Ventoso, I., R. Blanco, C. Perales, and L. Carrasco.** 2001. HIV-1 protease cleaves eukaryotic initiation factor 4G and inhibits cap-dependent translation. Proc. Natl. Acad. Sci. USA **98:**12966–12971.
50. **Yang, Z. R., R. Thomson, T. C. Hodgman, J. Dry, A. K. Doyle, A. Narayanan, and X. Wu.** 2003. Searching for discrimination rules in protease proteolytic cleavage activity using genetic programming with a min-max scoring function. Biosystems **72:**159–176.
51. **Yang, Z. R., and K. C. Chou.** 2004. Bio-support vector machines for computational proteomics. Bioinformatics **20:**735–741.
52. **Yang, Z. R., J. Qiu, and A. Dalby.** 2004. Mining HIV protease cleavage data using genetic programming with a sum-product function. Bioinformatics **20:**3398–3405.
53. **You, L., and T. Rögnvaldsson.** 2004. HIV protease cleavage specificity: model complexity, prediction accuracy and biological rule extraction. Poster Abstract at the 12th International Conference on Intelligent Systems for Molecular Biology and the 3rd European Conference on Computational Biology, 31 July to 4 August 2004, Glasgow, Scotland. [Online.] http://www.iscb.org/ismb2004/posters/liwen.youATide.hh.se_496.html.
54. **Zhang, Y.-M., H. Imamichi, T. Imamichi, H. C. Lane, J. Falloon, M. B. Vasudevachari, and N. P. Salzman.** 1997. Drug resistance during indinavir therapy is caused by mutations in the protease gene and in its Gag substrate cleavage sites. J. Virol. **71:**6662–6670.

Downloaded from http://jvi.asm.org/ on March 2, 2015 by HOGSKOLAN I HALMSTAD

JOURNAL OF VIROLOGY, Apr. 2006, p. 4205 Vol. 80, No. 8
0022-538X/06/$08.00+0 doi:10.1128/JVI.80.8.4205.2006

AUTHOR'S CORRECTION

Comprehensive Bioinformatic Analysis of the Specificity of Human Immunodeficiency Virus Type 1 Protease

Liwen You, Daniel Garwicz, and Thorsteinn Rögnvaldsson

School of Information Science, Computer and Electrical Engineering, Halmstad University, Halmstad, Sweden, and Division of Hematology and Transfusion Medicine, Department of Laboratory Medicine, Lund University, Lund, Sweden, and Division of Molecular Toxicology, Institute of Environmental Medicine, Karolinska Institutet, Stockholm, Sweden

Volume 79, number 19, p. 12477–12486, 2005. Page 12480: Figure 3 and its legend should appear as shown below. The out-of sample prediction performance for the Gaussian support vector machine (GSVM) algorithm was overestimated due to a computational mistake. As a result, the GSVM algorithm with hydrophobicity and size coding does not outperform the linear algorithms with sparse orthogonal coding. However, the two physicochemical parameters hydrophobicity and size are still the best pair of properties for predicting cleavage by the HIV-1 protease. As previously stated, there is no statistically significant difference (at the 95% level) in prediction performance between the best method using sparse orthogonal coding and the GSVM model with property coding. None of the other results or conclusions in the original paper are affected by the computational mistake.

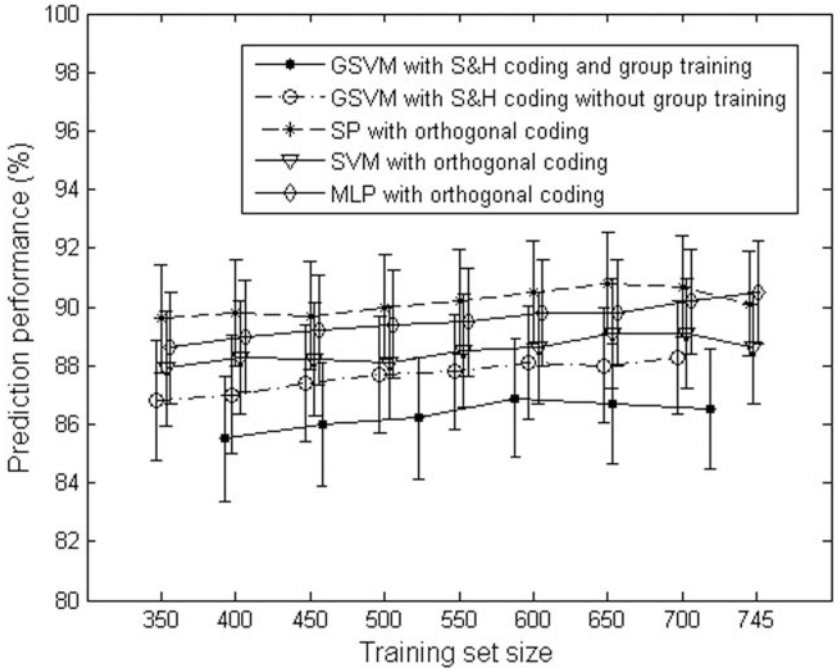

FIG. 3. The best predictors' out-of-sample performances, estimated using cross-validation. There is no statistically significant difference (at the 95% level) between the best linear and the best nonlinear predictors. The two bottom curves are both for property-coded data, but the upper one represents the case when care is not taken to avoid sequence bias in the testing (shown here to illustrate the importance of avoiding such bias and overly optimistic results). Here S denotes small property and H denotes hydrophobicity.