Predicting Human Immunodeficiency Virus Protease Cleavage Sites in Proteins by a Discriminant Function Method

Kuo-Chen Chou, Alfredo G. Tomasselli, Ilene M. Reardon, and Robert L. Heinrikson
Pharmacia & Upjohn Laboratories, Kalamazoo, Michigan 49001-4940

ABSTRACT **Based on the sequence-coupled (Markov chain) model and vector-projection principle, a discriminant function method is proposed to predict sites in protein substrates that should be susceptible to cleavage by the HIV-1 protease. The discriminant function is defined by $\Delta = \Phi^+ - \Phi^-$, where Φ^+ and Φ^- are the cleavable and noncleavable attributes for a given peptide, and they can be derived from two complementary sets of peptides, S^+ and S^-, known to be cleavable and noncleavable, respectively, by the enzyme. The rate of correct prediction by the method for the 62 cleavable peptides and 239 noncleavable peptides in the training set are 100 and 96.7%, respectively. Application of the method to the 55 sequences which are outside the training set and known to be cleaved by the HIV-1 protease accurately predicted 100% of the peptides as substrates of the enzyme. The method also predicted all but one of the sites hydrolyzed by the protease in native HIV-1 and HIV-2 reverse transcriptases, where the HIV-1 protease discriminates between nearly identical sequences in a very subtle fashion. Finally, the algorithm predicts correctly all of the HIV-1 protease processing sites in the native gag and gag/pol HIV-1 polyproteins, and all of the cleavage sites identified in denatured protease and reverse transcriptase. The new predictive algorithm provides a novel route toward understanding the specificity of this important therapeutic target.** © 1996 Wiley-Liss, Inc.

Key words: Markov chain, vector projection, forbidden rule, accessibility

INTRODUCTION

Interest in the protease from human immunodeficiency virus (HIV), the causative agent in acquired immunodeficiency syndrome (AIDS), derives from a number of considerations. The most obvious is as a target for design of drugs that might be beneficial in AIDS therapy. The HIV protease is essential for processing of the viral polyproteins in the final maturation phase of the life cycle leading to infectious virus, and inhibitors of the protease have been shown to block viral proliferation. Another possible basis for interest in the HIV protease is as a tool in protein chemistry. The enzyme displays a complex specificity, but one which can be directed toward highly variable, but rationally designed peptide sequences. Thus, it may find application as a means for excision of desired protein products from fusion protein constructs, where fidelity of the protein sequence at the termini of the product is essential. Finally, the HIV protease is something of a curiosity among proteolytic enzymes in that it is an obligate homodimer. Mechanistically, it is an aspartyl protease and typical members of this set such as pepsin and renin exist as single polypeptide chains in which the two similarly folded halves of the molecule each contribute an aspartyl residue to the active site. In the case of the retroviral protease homodimer, each monomer contributes a single Asp residue to the catalytic site in the symmetrical enzyme. The architecture of the HIV protease, therefore, provides an interesting example of symmetry upon which to base strategies for design of inhibitory molecules and considerable progress has been made in this area of research.

Underlying all of these important features of the HIV protease is the question of enzyme specificity. This subject has been dealt with by a number of experimental approaches. The first was to focus on the physiological function of the protease. The HIV, and other retroviral proteases, have evolved to carry out a very specific function in processing 7 or 8 bonds in the gag/ or gag/pol viral polyproteins to yield itself, and the various structural proteins and enzymes of the mature infectious virus. Although these processing sites are different, there are similarities among them which served as the basis for early speculation about specificity.[1] A later approach involved extensive studies of nonviral protein substrates. This revealed that the HIV protease

Received May 19, 1995; revision accepted July 28, 1995.

Address reprint requests to Kuo-Chen Chou, Computer-Aided Drug Discovery, Pharmacia & Upjohn Laboratories, Kalamazoo, MI 49001-4940.

has a broad and not easily understood specificity toward a wide variety of substrates.[2–4] These, and other findings suggested that the minimal size of a peptide that could be cleaved by this enzyme was 7 residues and this, in turn, led to the idea that the HIV-1 protease recognizes an extended amino acid sequence best represented by an octapeptide. Accordingly, many kinetic studies were carried out in which select octapeptide substrates were varied systematically and tested as substrates of the enzyme.[4–8] Particular amino acids appeared to be forbidden at defined positions from P_4 through $P_{4'}$,[4,9] although the notion emerged that the HIV-1 protease can hydrolyze almost any peptide bond, given accessibility to an extended, acceptable sequence encompassing 4 residues on either side of the scissile bond. At the time of this writing, more than 60 cleavable sequences have been identified in protein substrates,[2] and the list comprises more than 100 when one includes systematically varied peptide substrates.

Paralleling these experimental strategies, was an effort to devise algorithms for predicting sites of cleavage by the HIV proteases. The first such algorithm was published by Poorman et al.[10] and provided means for calculation of a probability function h, based upon the amino acid sequence of the octapeptide corresponding to P_4 through $P_{4'}$. This method has proven to be useful and easy to apply, but suffers from a number of shortcomings. The h function is a multiplication of $n_{i,j}$ or $s_{i,j}$, the frequency of the jth ($j = 1, 2, \ldots, 20$) amino acid occurring at the ith ($i = 4, 3, 2, 1, 1', 2', 3', 4'$) subsite for a given training set of cleavable peptides. When any of these frequencies was zero, some value, such as 0.25 or 0.5, must be assigned to it, thus inevitably introducing a measure of arbitrariness. The problem caused by such an arbitrary assignment is particularly serious when the data in the training set are limited. It should also be noted that in the h function method no clear procedure was described in determining the "cutoff value," a critical quantity in predicting the cleavability of an oligopeptide. The ambiguous treatment of such a critical quantity might introduce even more arbitrariness. Finally, in calculating the h function, the probability of an amino acid occurring in each of the eight specificity subsites was treated as a completely independent event. In other words, no coupling effect from neighboring amino acids along the peptide sequence was taken into account. Obviously, this will certainly affect the accuracy of prediction.

To deal with the first and second problems, the *vector projection approach*[11] and the *correlation-angle approach*[12] were proposed, and the percentage of correct prediction was increased somewhat accordingly. However, neither of these two methods incorporates the sequence coupling effect. Subsequently, a *Markov chain model*[13] and an *alternate-subsite-coupled model*[14] were proposed in order to deal with the third problem. Although both models make allowance for incorporating the coupling effect along the subsites of HIV protease through either a Markov chain mechanism or an alternate-subsite interaction mechanism, they cannot avoid assigning arbitrary values for the parameters of some subsites, especially in the case of limited number of training data. To improve the prediction method by taking into account all these three aspects, a *vectorized sequence-coupling model*[15] was proposed. In this model, the arbitrary assignment for insufficient experimental data is avoided by a vectorization approach, the "cutoff value" or "threshold value" derived by an optimization procedure, and the coupling effect reflected by a conditional probability matrix.

In this article, a new predictive algorithm, the discriminant function method, is proposed. The new method possesses all of the advantages of the vectorized sequence-coupling model and, in addition, the labor for deriving the cutoff value by the optimization procedure can be avoided because there is no need whatsoever to introduce such a quantity in the new method. Moreover, new experimental data have been incorporated into the current algorithm and some rules regarding "forbidden" amino acids have been observed. The algorithm has been applied to a number of practical examples wherein the patterns of hydrolysis of selective proteins by the HIV-1 protease have been determined experimentally.

METHOD

HIV proteases have extended substrate binding regions in which usually as many as eight consecutive amino acid moieties of the polypeptide substrate are in contact with the active-site cleft (Fig. 1). Because we deal with a probability function P, we use the symbol R rather than P as in the Schecter–Berger notation used originally by Schecter and Berger[16] to refer to amino acid residues at various positions in the octapeptide substrates. In studying the specificity of HIV protease, peptides can be classified into two categories: the positive set and negative set. The positive set, denoted by S^+, consists of peptides which are cleavable by the enzyme, while the negative set, S^-, consists of noncleavable ones.

Given an octapeptide, its attribute to the positive set S^+ or the negative set S^- can be formulated by an 8-D (dimension) vector. If the amino acid residue at each of the eight subsites can be treated as an independent element, i.e., there is no coupling at all among these subsites, then its attribute to the positive set S^+ and that to the negative set S^- can be expressed respectively as

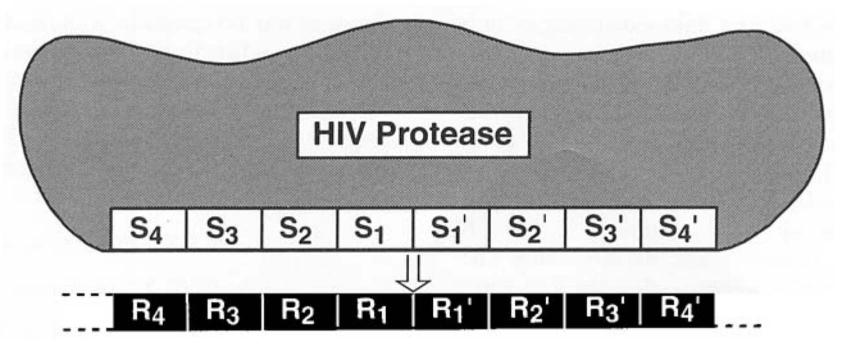

Fig. 1. Schematic representation of the enzyme–substrate complex. The active site of the enzyme (HIV-protease) is composed of eight "subsites," S_4, S_3, S_2, S_1, $S_{1'}$, $S_{2'}$, $S_{3'}$, $S_{4'}$, and their counterparts in a given protein extend to an octapeptide region, sequentially symbolized by R_4, R_3, R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, $R_{4'}$, respectively. The scissile bond is located between the subsites R_1 and $R_{1'}$. We use the symbol R rather P as introduced originally by Schechter and Berger[16] to avoid confusion with P for probability.

$$\mathbf{V}_0^+(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) = \begin{bmatrix} P_4^+(\mathrm{X}_4) \\ P_3^+(\mathrm{X}_3) \\ P_2^+(\mathrm{X}_2) \\ P_1^+(\mathrm{X}_1) \\ P_{1'}^+(\mathrm{X}_{1'}) \\ P_{2'}^+(\mathrm{X}_{2'}) \\ P_{3'}^+(\mathrm{X}_{3'}) \\ P_{4'}^+(\mathrm{X}_{4'}) \end{bmatrix} \qquad (1\mathrm{a})$$

and

$$\mathbf{V}_0^-(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) = \begin{bmatrix} P_4^-(\mathrm{X}_4) \\ P_3^-(\mathrm{X}_3) \\ P_2^-(\mathrm{X}_2) \\ P_1^-(\mathrm{X}_1) \\ P_{1'}^-(\mathrm{X}_{1'}) \\ P_{2'}^-(\mathrm{X}_{2'}) \\ P_{3'}^-(\mathrm{X}_{3'}) \\ P_{4'}^-(\mathrm{X}_{4'}) \end{bmatrix} \qquad (1\mathrm{b})$$

where P_i^+ (X_i) $(i = 4, 3, 2, 1, 1', 2', 3', 4')$ is the probability of amino acid X_i occurring at subsite R_i in the positive set S^+, and its value can be derived from a set of training data consisting of only the peptides known to be cleavable by HIV protease. P_i^- (X_i) in Eq. (1b) has the same meaning as P_i^+ (X_i) of Eq. (1a) except that it is associated with the negative set S^-, and its value should be derived from a set of training data consisting of only those peptides known to be not cleavable by the enzyme. The subscript 0 of **V** marks that the components of the 8-D vector is formed by eight independent probabilities in which no coupling effect between subunits is included, as shown by the right side of Eq. (1). These independent probabilities actually correspond to the zero-order coupled terms.

However, if the coupling effect of a residue with those adjacent to it (Fig. 1) must be taken into account, then the matrix elements in Eq. (1) should be modified according to the first-order Markov chain theory,[32] i.e., substituted by the first-order conditional probabilities. For consistency, the corresponding vector symbol $\mathbf{V}_0$ in Eq. (1) should also be changed to $\mathbf{V}_1$, or simply to **V**. Thus, instead of Eq. (1a) we should have

$$\mathbf{V}^+(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) = \begin{bmatrix} P_4^+(\mathrm{X}_4) \\ P_3^+(\mathrm{X}_3|\mathrm{X}_4) \\ P_2^+(\mathrm{X}_2|\mathrm{X}_3) \\ P_1^+(\mathrm{X}_1|\mathrm{X}_2) \\ P_{1'}^+(\mathrm{X}_{1'}|\mathrm{X}_1) \\ P_{2'}^+(\mathrm{X}_{2'}|\mathrm{X}_{1'}) \\ P_{3'}^+(\mathrm{X}_{3'}|\mathrm{X}_{2'}) \\ P_{4'}^+(\mathrm{X}_{4'}|\mathrm{X}_{3'}) \end{bmatrix} \qquad (2\mathrm{a})$$

where $\mathrm{P}_4^+(\mathrm{X}_4)$ is the same as in Eq. (1a), i.e., the probability of amino acid X_4 occurring at subsite position R_4 in the positive set S^+ and it is independent of the other subsites because R_4 is located at the first position of the eight-subsite sequence (Fig. 1). $\mathrm{P}_3^+(\mathrm{X}_3|\mathrm{X}_4)$ is the probability of amino acid X_3 occurring at the subsite R_3, given that X_4 has occurred at position R_4; $\mathrm{P}_2^+(\mathrm{X}_2|\mathrm{X}_3)$ is the probability of amino acid X_2 occurring at the subsite R_2, given that X_3 has occurred at position R_3, and so forth. Similarly, instead of Eq. (1b) we should have

$$\mathbf{V}^-(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) = \begin{bmatrix} P_4^-(\mathrm{X}_4) \\ P_3^-(\mathrm{X}_3|\mathrm{X}_4) \\ P_2^-(\mathrm{X}_2|\mathrm{X}_3) \\ P_1^-(\mathrm{X}_1|\mathrm{X}_2) \\ P_{1'}^-(\mathrm{X}_{1'}|\mathrm{X}_1) \\ P_{2'}^-(\mathrm{X}_{2'}|\mathrm{X}_{1'}) \\ P_{3'}^-(\mathrm{X}_{3'}|\mathrm{X}_{2'}) \\ P_{4'}^-(\mathrm{X}_{4'}|\mathrm{X}_{3'}) \end{bmatrix} \qquad (2\mathrm{b})$$

where all the elements have the same meaning as those of Eq. (2a) except that they are associated with the negative set S^- and their values should be derived from a set of training data consisting of only the noncleavable peptides.

Generally speaking, if the coupling effects of the l (l = 2,3, . . .) closest neighboring amino acid residues need to be considered, then Eq. (1) should be modified according to the lth-order Markov chain theory, i.e., the vector symbol $\mathbf{V}_0$ should be replaced by $\mathbf{V}_l$ and the corresponding matrix elements by the lth-order conditional probabilities. As one could surmise, the analysis of a higher-order Markov chain would be much more complicated. Therefore, the treatment in this paper is confined to the first-order Markov chain, i.e., only the first-order sequence-coupling effect is taken into account, as formulated by Eq. (2).

Now in the 8-D space, let us define an ideal cleavability-positive vector, Λ^+, each of whose eight components λ_i^+ (i = 4, 3, 2, 1, 1′, 2′, 3′, 4′) is the upper limit of the corresponding matrix element in Eq. (2a). Theoretically, the upper limit is 1, meaning that Λ^+ would be the vector for a *hypothetical, idealized* peptide which would be the only cleavable peptide for HIV protease. Therefore, for such an ideal cleavability-positive vector Λ^+, all of its components are equal to 1. The similarity in the cleavability-positive attribute between a given octapeptide and the idealized cleavable peptide can be expressed in terms of the projection of $\mathbf{V}^+$ on Λ^+. The larger the projection, the higher the similarity, and hence the closer the peptide to the cleavability-positive set. This is the so-called maximum-vector-projection principle, or maximum-correlation-coefficient principle,[17] which has proved to be quite successful when used to predict the structural class of a protein from its amino acid composition. Accordingly, the attribute function of a given octapeptide to the cleavability-positive set can be formulated by

$$\begin{aligned}\Psi^+(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) &= \mathbf{V}^+\cdot\Lambda^+\\ &= P_4^+(X_4) + P_3^+(X_3|X_4) + P_2^+(X_2|X_3) + P_1^+(X_1|X_2)\\ &\quad + P_{1'}^+(X_{1'}|X_1) + P_{2'}^+(X_{2'}|X_{1'}) + P_{3'}^+(X_{3'}|X_{2'}) + P_{4'}^+(X_{4'}|X_{3'})\end{aligned} \tag{3a}$$

On the other hand, we can also in the 8-D space define an ideal cleavability-negative vector, Λ^-, each of whose eight components λ_i^- (i = 4, 3, 2, 1, 1′, 2′, 3′, 4′) is the upper limit of the corresponding matrix element in Eq. (2b). Theoretically, the upper limit is also 1, meaning that Λ^- would be the vector for a *hypothetical, idealized* peptide which would be the only noncleavable peptide for the enzyme. Thus, it follows according to the similar rationale that the attribute function of a given octapeptide to the cleavability-negative set can be formulated by

$$\begin{aligned}\Psi^-(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) &= \mathbf{V}^-\cdot\Lambda^-\\ &= P_4^-(X_4) + P_3^-(X_3|X_4) + P_2^-(X_2|X_3) + P_1^-(X_1|X_2)\\ &\quad + P_{1'}^-(X_{1'}|X_1) + P_{2'}^-(X_{2'}|X_{1'}) + P_{3'}^-(X_{3'}|X_{2'}) + P_{4'}^-(X_{4'}|X_{3'})\end{aligned} \tag{3b}$$

For a given octapeptide $X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}$, if its attribute function to the cleavability-positive set is greater than that to the cleavability-negative set, i.e., $\Psi^+ > \Psi^-$, then the peptide is predicted to be a cleavable one; otherwise, it is predicted to be a noncleavable one. Define a discriminant function Δ given by

$$\begin{aligned}&\Delta(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'})\\ &\quad = \Psi^+(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'})\\ &\quad -\Psi^-(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) + \Re\end{aligned} \tag{4}$$

where $\Re$ is a modified factor associated with some special empirical rules as will be described later [see Eq. (6)]. If no empirical rules are incorporated, one may just set $\Re = 0$. Thus, the criterion for predicting the cleavability of a peptide can be formulated in terms of its discriminant function Δ as follows:

$$\begin{cases}\text{A peptide is cleavable by HIV protease} & \text{if its } \Delta > 0\\ \text{A peptide is non-cleavable by HIV-protease,} & \text{otherwise}\end{cases} \tag{5}$$

If, occasionally, the peptide to deal with is shorter than an octapeptide, such as for the case of a heptapeptide,[10] one can simply set zero for the probability term of the absent residue. For example, if the peptide to be predicted is $X_3X_2X_1X_{1'}X_{2'}X_{3'}$ $X_{4'}$, then in Eqs. (2)–(3), one should substitute zero for $P^+(X_4)$ and $P^-(X_4)$, because there is no residue at the subsites R_4 for the peptide concerned. Also, substitute $P^+(X_3)$ for $P_3^+(X_3|X_4)$ and $P^-(X_3)$ for $P_3^-(X_3|X_4)$ since in this case any coupling associated with subsite R_4 would vanish.

The formulation given above can be used to predict the cleavage sites by both HIV-1 and HIV-2 proteases. However, for the case of HIV-1 protease, the positive and negative training sets S^+ and S^- should consist of the peptides associated with HIV-1 protease; while for the case of HIV-2 protease, the corresponding S^+ and S^- sets should be associated with HIV-2 protease.

It has been observed[18] that some residues are not tolerated at particular subsites for the cleavable peptides by HIV-1 protease. For example, Lys residues appear to be forbidden anywhere from R_2 through $R_{2'}$. Since Lys is an abundant amino acid, its prohibition in this stretch of sequence should have an important impact on the algorithm. To incorporate this into the algorithm, the modified factor $\Re$ in Eq. (4) for HIV-1 protease should be given as follows:

$$\Re = \begin{cases}\Re_K, & \text{if K is at subsite } R_2, R_1, R_{1'}, \text{ or } R_{2'},\\ 0, & \text{otherwise}\end{cases} \tag{6}$$

where $\Re_K$ can be any large negative number as long as it can lead to $\Delta < 0$ [Eq. (4)] when the intolerable residue K occurs at any of the forbidden subsites. In this paper $\Re_K = -3$.

RESULTS AND DISCUSSION

In order to calculate the attribute functions Ψ^+ for any octapeptide, we have to first find $P_i^+(\mathrm{X})$ (i = 4, 3, 2, 1, 1′, 2′, 3′, 4′), as well as $P_3^+(\mathrm{X}_3|\mathrm{X}_4)$, $P_2^+(\mathrm{X}_2|\mathrm{X}_3)$, $P_1^+(\mathrm{X}_1|\mathrm{X}_2)$, $P_{1'}^+(\mathrm{X}_{1'}|\mathrm{X}_1)$, $P_{2'}^+(\mathrm{X}_{2'}|\mathrm{X}_{1'})$, $P_{3'}^+(\mathrm{X}_{3'}|\mathrm{X}_{2'})$, $P_{4'}^+(\mathrm{X}_{4'}|\mathrm{X}_{3'})$. These we derive from a positive training set consisting of 62 cleavable peptides (Table I) taken from Table 3 of Tomasselli et al.[2] by leaving out two modified pepdides.

In order to calculate the attribute functions Ψ^- for any octapeptide, we have to find $P_i^-(\mathrm{X})$(i = 4, 3, 2, 1, 1′, 2′, 3′, 4′), as well as P_3^- $(\mathrm{X}_3|\mathrm{X}_4)$, $P_2^-(\mathrm{X}_2|\mathrm{X}_3)$, $P_1^-(\mathrm{X}_1|\mathrm{X}_2)$, $P_0^-(\mathrm{X}_0|\mathrm{X}_1)$, $P_{1'}^-(\mathrm{X}_{1'}|\mathrm{X}_0)$, $P_{2'}^-(\mathrm{X}_{2'}|\mathrm{X}_{1'})$, $P_{3'}^-(\mathrm{X}_{3'}|\mathrm{X}_{2'})$, $P_{4'}^-(\mathrm{X}_{4'}|\mathrm{X}_{3'})$. These we derive from a negative training set consisting of 239 noncleavable octapeptides (Table II), of which 122 (=129−7) are extracted from hen egg lysozyme and 117 (=124−7) from bovine pancreatic ribonuclease, since neither of the two proteins had shown any probable cleavage sites even if they are completely denatured to make any part of them accessible to the active site of HIV-1 protease.[10]

The computations were carried out on an IBM 3090/400J computer at Upjohn Laboratories. Listed in Table III are the probability values for $P_i^+(\mathrm{X})$ (i = 4, 3, 2, 1, 1′, 2′, 3′, 4′) derived from the 62 cleavable peptides (Table I), while the relevant conditional probabilities, which contain 20 × 20 × 7 = 2800 data, are given in Appendix A. Listed in Table IV are the probability values for $P_i^-(\mathrm{X})$ (i = 4, 3, 2, 1, 1′, 2′, 3′, 4′) derived from the 239 noncleavable peptides (Table II), while the relevant conditional probabilities, which also contain 20 × 20 × 7 = 2800 data, are given in Appendix B.

Based on the data in Tables III and IV and Appendices A and B, the discriminant function Δ for any given peptide $\mathrm{X}_4\mathrm{X}_3\mathrm{X}_2\mathrm{X}_1\mathrm{X}_{1'}\mathrm{X}_{2'}\mathrm{X}_{3'}\mathrm{X}_{4'}$ can be calculated by means of Eqs. (3) and (4), and its cleavability of the peptide can then be predicted according to Eq. (5).

Predicted Results for the Training Data

The predicted results by the new method for the 62 peptides in the cleavability-positive training set, together with those by the h function method, are given in Table I. As can be seen there, all these peptides have $\Delta > 0$, indicating that the percentage of correct prediction for the cleavability-positive training set is 62/62 = 100%. According to the h function method, 22 of them have $h < 0.13$, meaning that the percentage of correct prediction is only 40/62 = 64.5%. The predicted results for the cleavability-negative set are given in Table II, from which we can see that the percentage of correct prediction for the 239 noncleavable peptides is 231/239 = 96.7%. Accordingly, the percentages of correct prediction by the new method are very high for both the cleavable and noncleavable peptides in the training set database. It has been observed that either Ile or Val residue occurring at R_1 will significantly reduce the rate of cleavability,[18] causing the corresponding peptide to become a noncleavable one. Interestingly, 19 such cases are found in Table II, and they are GNWV-CAAK, DYGI-LQIN, ILQI-NSRW, LCNI-PCSA, SSDI-TASV, TASV-NCAK, AKKI-VSDG, KKIV-SDGN, NAWV-AWRN, GTDV-QAWI, QAWI-RGCR, CKPV-NTFV, NTFV-HESL, LADV-QAVC, VQAV-CSQK, QKNV-ACKN, HIIV-ACEG, NPYV-PVHF, and YVPV-HFDA. All these peptides, except one (i.e., DYGI-LQIN), are correctly predicted to be noncleavable by HIV-1 protease. This indicates that the negative effect caused by Ile or Val residue at R_1 has been automatically reflected by the current algorithm.

Predicted Results for the Testing Data

Listed in Table V are the predicted results for a set of testing data which are outside the training set data. The testing set consists of 55 peptides known to be cleavable by HIV-1 protease. It can be seen from Table V that all these peptides have a value of $\Delta > 0$, indicating that they are all correctly predicted to be cleavable. If predicted by the h-function method, 7 peptides are incorrectly predicted to be noncleavable. Therefore for the 55 testing peptides, the percentage of correct prediction by the current method is 12.7% higher than the h-function method.

Application of the Algorithm

The use of algorithms to identify potential sites of proteolysis in folded, native proteins is, of course, limited by the caveat that particular bonds may be perfectly acceptable but inaccessible to the enzyme. Even in denatured protein substrates it is not always clear that there does not remain some element of secondary or supersecondary structure that limits needed accessibility of the protease to the predicted site. Retroviral proteases such as the HIV-1 enzyme have evolved to process only a few select bonds in the gag/pol polyprotein precursors, and this would suggest that the individual proteins packaged into the polyprotein format are folded and protected at internal sites, with accessibility limited to the segments of structure which connect them one to the other. Predictive algorithms applied to the sequence per se should, therefore, overpredict.

In the case of the HIV-1 protease, which appears to require at least 7 amino acids in peptide substrates,[25] there is another point of concern. One may well predict 2 or 3 sites of cleavage within a given sequence of, say, a dozen amino acids. In fact, because the specificity of the protease is cumulative,[10] and fairly independent of the nature of P_1 and $\mathrm{P}_{1'}$ amino acids, it stands to reason that particular regions of sequence may have more than one bond susceptible for hydrolysis. However, if one of the sites

TABLE I. The 62 Cleavable Peptides by HIV-1 Protease*

Peptide sequence and cleavage site											
R_4	R_3	R_2	R_1	⇓	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	Δ[†]	$h - 0.13$[‡]	Protein
T	Q	I	M	⇓	F	E	T	F	1.22	0.84	Actin
G	Q	V	N	⇓	Y	E	E	F	1.24	0.83	Calmodulin
P	F	I	F	⇓	E	E	E	P	2.10	0.83	pro-IL1-β
S	F	N	F	⇓	P	Q	I	T	1.38	0.79	*pol*
D	T	V	L	⇓	E	E	M	S	2.03	0.77	Autolysis
A	R	V	L	⇓	A	E	A	M	1.77	0.76	*gag*
A	E	E	L	⇓	A	E	I	F	2.30	0.76	Troponin C
S	L	N	L	⇓	R	E	T	N	1.17	0.74	Vimentin
A	T	I	M	⇓	M	Q	R	G	1.47	0.69	*gag*
A	E	C	F	⇓	R	I	F	D	2.68	0.69	Troponin C
D	Q	I	L	⇓	I	E	I	C	1.45	0.68	Autolysis
D	D	L	F	⇓	F	E	A	D	1.09	0.64	pro-IL1-β
Y	E	E	F	⇓	V	Q	M	M	1.89	0.62	Calmodulin
P	I	V	G	⇓	A	E	T	F	1.79	0.62	*pol*
T	L	N	F	⇓	P	I	S	P	2.17	0.61	*pol*
R	E	A	F	⇓	R	V	F	D	1.27	0.59	Calmodulin
A	E	T	F	⇓	Y	V	D	K	1.90	0.55	*pol*
A	Q	T	F	⇓	Y	V	N	L	1.51	0.45	*pol*
P	T	L	L	⇓	T	E	A	P	1.89	0.44	Actin
S	F	I	G	⇓	M	E	S	A	1.43	0.40	Actin
D	A	I	N	⇓	T	E	F	K	2.22	0.34	Vimentin
Q	I	T	L	⇓	W	Q	R	P	1.75	0.33	Autolysis
E	L	E	F	⇓	P	E	G	G	1.90	0.33	PE664E
	A	N	L	⇓	A	E	E	A	1.64	0.26	PE40
S	Q	N	Y	⇓	P	I	V	Q	1.35	0.25	*gag*
P	G	N	F	⇓	L	Q	S	R	1.23	0.25	*gag*
K	L	V	F	⇓	F	A	E		1.46	0.24	AAP[§]
G	D	A	L	⇓	L	E	R	N	1.03	0.19	PE40
K	E	L	Y	⇓	P	L	T	S	1.21	0.15	*gag*
R	Q	A	N	⇓	F	L	G	K	1.42	0.08	*gag*
S	R	S	L	⇓	Y	A	S	S	1.18	0.07	Vimentin
A	E	A	M	⇓	S	Q	V	T	2.28	0.04	*gag*
R	K	I	L	⇓	F	L	D	G	1.79	−0.01	*pol*
G	S	H	L	⇓	V	E	A	L	2.62	−0.03	Insulin
G	G	V	Y	⇓	A	T	R	S	1.58	−0.04	Vimentin
F	R	S	G	⇓	V	E	T	T	2.89	−0.04	*gag*
V	E	V	A	⇓	E	E	E	E	2.58	−0.05	AAP[§]
L	P	V	N	⇓	G	E	F	S	2.69	−0.05	AAP[§]
E	T	T	A	⇓	L	V	C	D	1.65	−0.10	Actin
H	L	V	E	⇓	A	L	Y	L	2.59	−0.11	Insulin**
H	Y	G	F	⇓	P	T	Y	G	3.52	−0.13	NF-κB[††]
D	S	A	D	⇓	A	E	E	D	2.68	−0.11	AAP[§]
G	W	I	L	⇓	G	E	H	G	2.88	−0.08	LDH[‡‡]
G	W	I	L	⇓	A	E	H	G	2.72	0.10	LDH
Q	A	I	Y	⇓	L	A	L	Q	1.70	−0.13	*pol*[§§]
E	K	V	Y	⇓	L	A	W	V	1.98	−0.13	*pol*
V	E	I	C	⇓	T	E	M	E	3.55	−0.06	*pol****
T	Q	D	F	⇓	W	E	V	Q	2.09	−0.02	*pol*
L	W	M	G	⇓	Y	E	L	H	2.56	−0.13	*pol*
G	D	A	Y	⇓	F	S	V	P	2.47	−0.12	*pol*
E	L	E	L	⇓	A	E	N	R	2.21	−0.02	*pol*
S	K	D	L	⇓	I	A	E	I	1.86	−0.13	*pol*
L	E	V	N	⇓	I	V	T	D	0.93	−0.03	*pol*
G	G	N	Y	⇓	P	V	Q	H	1.56	−0.12	*gag*[†††]
A	R	L	M	⇓	A	E	A	L	2.11	0.33	*gag*
P	F	A	A	⇓	A	Q	Q	R	1.35	−0.12	*gag*
P	R	N	F	⇓	P	V	A	Q	0.96	0.49	*gag*
G	L	A	A	⇓	P	Q	F	S	0.99	0.15	*gag/pol*
S	L	N	L	⇓	P	V	A	K	0.93	0.39	*pol*
A	E	T	F	⇓	Y	T	D	G	1.88	0.22	*pol*
R	Q	V	L	⇓	F	L	E	K	1.82	0.65	*pol*
Q	M	I	F	⇓	E	E	H	G	3.03	0.05	Fibronectin[‡‡‡]

*Extracted from Table 3 of Tomasselli et al.[2] Note that listed here are 62 rather than 64 peptides as in Table 3 of Tomasselli et al.[2] since two of them were chemically modified and should not be included here.

[†]Δ is the criterion used in this paper for predicting whether an oligopeptide can be cleaved by HIV-1 protease: an oligopeptide can be cleaved when its $\Delta \geq 0$; otherwise, it cannot be cleaved. The values of Δ were calculated according to Eqs. (3)–(4).

[‡]h is the criterion used in the h function method[10] to predict whether an octapeptide can be cleaved by HIV-1 protease: an oligopeptide can be cleaved when its $h \geq 0.13$; otherwise, it cannot be cleaved.

[§]Alzheimer amyloid protein.

**All entires to this point were referenced in Poorman et al.[10]

[††]Riviere et. al.[19]

[‡‡]Tomaszek et al.[20]

[§§]The following two entires are from Chattopadhyay et al.[21]

***The following seven entires are from Tomasselli et al.[2]

[†††]The following eight entires are from Tözsér et al.[9]

[‡‡‡]Oswald and von der Helm.[22]

TABLE II. The 239 Noncleavable Peptides by HIV-1 Protease*

Peptide sequence†									
R_4	R_3	R_2	R_1	⋈	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	$\Delta^\ddagger$
K	V	F	G	⋈	R	C	E	L	−1.02
V	F	G	R	⋈	C	E	L	A	−1.14
F	G	R	C	⋈	E	L	A	A	−1.29
G	R	C	E	⋈	L	A	A	A	−0.78
R	C	E	L	⋈	A	A	A	M	−0.28
C	E	L	A	⋈	A	A	M	K	−1.09
E	L	A	A	⋈	A	M	K	R	−0.18
L	A	A	A	⋈	M	K	R	H	−4.00
A	A	A	M	⋈	K	R	H	G	−0.07
A	A	M	K	⋈	R	H	G	L	−1.00
A	M	K	R	⋈	H	G	L	D	−3.85
M	K	R	H	⋈	G	L	D	N	−0.77
K	R	H	G	⋈	L	D	N	Y	−0.72
R	H	G	L	⋈	D	N	Y	R	−0.78
H	G	L	D	⋈	N	Y	R	G	−0.60
G	L	D	N	⋈	Y	R	G	Y	−0.30
L	D	N	Y	⋈	R	G	Y	S	−0.60
D	N	Y	R	⋈	G	Y	S	L	−0.77
N	Y	R	G	⋈	Y	S	L	G	−0.61
Y	R	G	Y	⋈	S	L	G	N	−0.71
R	G	Y	S	⋈	L	G	N	W	−0.81
G	Y	S	L	⋈	G	N	W	V	0.56
Y	S	L	G	⋈	N	W	V	C	−1.18
S	L	G	N	⋈	W	V	C	A	−0.71
L	G	N	W	⋈	V	C	A	A	−0.55
G	N	W	V	⋈	C	A	A	K	−0.99
N	W	V	C	⋈	A	A	K	F	−1.34
W	V	C	A	⋈	A	K	F	E	−4.45
V	C	A	A	⋈	K	F	E	S	−1.39
C	A	A	K	⋈	F	E	S	N	−1.29
A	A	K	F	⋈	E	S	N	F	−4.23
A	K	F	E	⋈	S	N	F	N	−1.28
K	F	E	S	⋈	N	F	N	T	−1.34
F	E	S	N	⋈	F	N	T	Q	−1.07
E	S	N	F	⋈	N	T	Q	A	−0.78
S	N	F	N	⋈	T	Q	A	T	−0.81
N	F	N	T	⋈	Q	A	T	N	−0.75
F	N	T	Q	⋈	A	T	N	R	−0.51
N	T	Q	A	⋈	T	N	R	N	−1.03
T	Q	A	T	⋈	N	R	N	T	−0.43
Q	A	T	N	⋈	R	N	T	D	−0.80
A	T	N	R	⋈	N	T	D	G	0.00
T	N	R	N	⋈	T	D	G	S	−1.17
N	R	N	T	⋈	D	G	S	T	−1.25
R	N	T	D	⋈	G	S	T	D	−1.35
N	T	D	G	⋈	S	T	D	Y	−1.07
T	D	G	S	⋈	T	D	Y	G	−0.81
D	G	S	T	⋈	D	Y	G	I	−1.09
G	S	T	D	⋈	Y	G	I	L	−0.76
S	T	D	Y	⋈	G	I	L	Q	−0.34
T	D	Y	G	⋈	I	L	Q	I	−0.85
D	Y	G	I	⋈	L	Q	I	N	0.65
Y	G	I	L	⋈	Q	I	N	S	−0.35
G	I	L	Q	⋈	I	N	S	R	−0.34
I	L	Q	I	⋈	N	S	R	W	−0.72
L	Q	I	N	⋈	S	R	W	W	−0.39
Q	I	N	S	⋈	R	W	W	C	−0.46
I	N	S	R	⋈	W	W	C	N	−0.93

(continued)

TABLE II. The 239 Noncleavable Peptides by HIV-1 Protease* (*Continued*)

Peptide sequence†									
R_4	R_3	R_2	R_1	⋈	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	Δ‡
N	S	R	W	⋈	W	C	N	D	−0.91
S	R	W	W	⋈	C	N	D	G	−0.20
R	W	W	C	⋈	N	D	G	R	−0.96
W	W	C	N	⋈	D	G	R	T	−1.02
W	C	N	D	⋈	G	R	T	P	−0.95
C	N	D	G	⋈	R	T	P	G	−1.02
N	D	G	R	⋈	T	P	G	S	−1.07
D	G	R	T	⋈	P	G	S	R	−0.77
G	R	T	P	⋈	G	S	R	N	−0.73
R	T	P	G	⋈	S	R	N	L	−0.52
T	P	G	S	⋈	R	N	L	C	−1.07
P	G	S	R	⋈	N	L	C	N	−0.95
G	S	R	N	⋈	L	C	N	I	−0.87
S	R	N	L	⋈	C	N	I	P	−0.29
R	N	L	C	⋈	N	I	P	C	−1.01
N	L	C	N	⋈	I	P	C	S	−0.75
L	C	N	I	⋈	P	C	S	A	−0.57
C	N	I	P	⋈	C	S	A	L	−0.56
N	I	P	C	⋈	S	A	L	L	−0.61
I	P	C	S	⋈	A	L	L	S	−0.72
P	C	S	A	⋈	L	L	S	S	−0.28
C	S	A	L	⋈	L	S	S	D	−0.15
S	A	L	L	⋈	S	S	D	I	−0.46
A	L	L	S	⋈	S	D	I	T	−0.48
L	L	S	S	⋈	D	I	T	A	−0.87
L	S	S	D	⋈	I	T	A	S	−0.91
S	S	D	I	⋈	T	A	S	V	−0.67
S	D	I	T	⋈	A	S	V	N	0.27
D	I	T	A	⋈	S	V	N	C	−0.01
I	T	A	S	⋈	V	N	C	A	−1.01
T	A	S	V	⋈	N	C	A	K	−0.69
A	S	V	N	⋈	C	A	K	K	−0.52
S	V	N	C	⋈	A	K	K	I	−3.78
V	N	C	A	⋈	K	K	I	V	−4.00
N	C	A	K	⋈	K	I	V	S	−0.66
C	A	K	K	⋈	I	V	S	D	−3.58
A	K	K	I	⋈	V	S	D	G	−3.19
K	K	I	V	⋈	S	D	G	N	−0.68
K	I	V	S	⋈	D	G	N	G	−0.56
I	V	S	D	⋈	G	N	G	M	−1.08
V	S	D	G	⋈	N	G	M	N	−0.99
S	D	G	N	⋈	G	M	N	A	−0.72
D	G	N	G	⋈	M	N	A	W	−0.04
G	N	G	M	⋈	N	A	W	V	0.24
N	G	M	N	⋈	A	W	V	A	−1.16
G	M	N	A	⋈	W	V	A	W	−0.78
M	N	A	W	⋈	V	A	W	R	−1.03
N	A	W	V	⋈	A	W	R	N	−1.18
A	W	V	A	⋈	W	R	N	R	−0.75
W	V	A	W	⋈	R	N	R	C	−1.44
V	A	W	R	⋈	N	R	C	K	−1.27
A	W	R	N	⋈	R	C	K	G	−1.11
W	R	N	R	⋈	C	K	G	T	−3.97
R	N	R	C	⋈	K	G	T	D	−1.04
N	R	C	K	⋈	G	T	D	V	−0.84
R	C	K	G	⋈	T	D	V	Q	−3.63
C	K	G	T	⋈	D	V	Q	A	−1.25
K	G	T	D	⋈	V	Q	A	W	−0.95

(*continued*)

TABLE II. The 239 Noncleavable Peptides by HIV-1 Protease* (*Continued*)

Peptide sequence†									
R_4	R_3	R_2	R_1	⋈	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	Δ‡
G	T	D	V	⋈	Q	A	W	I	−1.06
T	D	V	Q	⋈	A	W	I	R	−1.41
D	V	Q	A	⋈	W	I	R	G	−0.99
V	Q	A	W	⋈	I	R	G	C	−1.08
Q	A	W	I	⋈	R	G	C	R	−0.82
A	W	I	R	⋈	G	C	R	L	−0.10
K	E	T	A	⋈	A	A	K	F	−0.14
E	T	A	A	⋈	A	K	F	E	−3.82
T	A	A	A	⋈	K	F	E	R	−1.25
A	A	A	K	⋈	F	E	R	Q	−1.06
A	A	K	F	⋈	E	R	Q	H	−3.61
A	K	F	E	⋈	R	Q	H	M	−1.16
K	F	E	R	⋈	Q	H	M	D	−1.26
F	E	R	Q	⋈	H	M	D	S	−1.17
E	R	Q	H	⋈	M	D	S	S	−0.68
R	Q	H	M	⋈	D	S	S	T	−0.45
Q	H	M	D	⋈	S	S	T	S	−0.83
H	M	D	S	⋈	S	T	S	A	−0.72
M	D	S	S	⋈	T	S	A	A	−0.95
D	S	S	T	⋈	S	A	A	S	−0.66
S	S	T	S	⋈	A	A	S	S	−0.58
S	T	S	A	⋈	A	S	S	S	−0.51
T	S	A	A	⋈	S	S	S	N	−0.26
S	A	A	S	⋈	S	S	N	Y	−1.00
A	A	S	S	⋈	S	N	Y	C	−0.99
A	S	S	S	⋈	N	Y	C	N	−0.93
S	S	S	N	⋈	Y	C	N	Q	−0.71
S	S	N	Y	⋈	C	N	Q	M	−0.59
S	N	Y	C	⋈	N	Q	M	M	−0.32
N	Y	C	N	⋈	Q	M	M	K	−1.14
Y	C	N	Q	⋈	M	M	K	S	−1.05
C	N	Q	M	⋈	M	K	S	R	−3.60
N	Q	M	M	⋈	K	S	R	N	−0.97
Q	M	M	K	⋈	S	R	N	L	−0.35
M	M	K	S	⋈	R	N	L	T	−4.20
M	K	S	R	⋈	N	L	T	K	−0.90
K	S	R	N	⋈	L	T	K	D	−0.86
S	R	N	L	⋈	T	K	D	R	−3.10
R	N	L	T	⋈	K	D	R	C	−0.85
N	L	T	K	⋈	D	R	C	K	−0.86
L	T	K	D	⋈	R	C	K	P	−3.79
T	K	D	R	⋈	C	K	P	V	−3.73
K	D	R	C	⋈	K	P	V	N	−1.18
D	R	C	K	⋈	P	V	N	T	−0.75
R	C	K	P	⋈	V	N	T	F	−3.83
C	K	P	V	⋈	N	T	F	V	−1.21
K	P	V	N	⋈	T	F	V	H	0.32
P	V	N	T	⋈	F	V	H	E	−1.15
V	N	T	F	⋈	V	H	E	S	−0.59
N	T	F	V	⋈	H	E	S	L	−1.22
T	F	V	H	⋈	E	S	L	A	−1.27
F	V	H	E	⋈	S	L	A	D	−1.09
V	H	E	S	⋈	L	A	D	V	−0.80
H	E	S	L	⋈	A	D	V	Q	0.09
E	S	L	A	⋈	D	V	Q	A	−1.20
S	L	A	D	⋈	V	Q	A	V	−0.25
L	A	D	V	⋈	Q	A	V	C	−1.19
A	D	V	Q	⋈	A	V	C	S	−0.94

(continued)

TABLE II. The 239 Noncleavable Peptides by HIV-1 Protease* (*Continued*)

Peptide sequence†									
R_4	R_3	R_2	R_1	⋈	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	$\Delta^‡$
D	V	Q	A	⋈	V	C	S	Q	−1.05
V	Q	A	V	⋈	C	S	Q	K	−0.92
Q	A	V	C	⋈	S	Q	K	N	0.32
A	V	C	S	⋈	Q	K	N	V	−3.60
V	C	S	Q	⋈	K	N	V	A	−0.87
C	S	Q	K	⋈	N	V	A	C	−0.60
S	Q	K	N	⋈	V	A	C	K	−3.69
Q	K	N	V	⋈	A	C	K	N	−0.93
K	N	V	A	⋈	C	K	N	G	−3.86
N	V	A	C	⋈	K	N	G	Q	−0.94
V	A	C	K	⋈	N	G	Q	T	−0.94
A	C	K	N	⋈	G	Q	T	N	−3.39
C	K	N	G	⋈	Q	T	N	C	−0.91
K	N	G	Q	⋈	T	N	C	Y	−0.73
N	G	Q	T	⋈	N	C	Y	Q	−0.78
G	Q	T	N	⋈	C	Y	Q	S	−0.37
Q	T	N	C	⋈	Y	Q	S	Y	−0.53
T	N	C	Y	⋈	Q	S	Y	S	−0.81
N	C	Y	Q	⋈	S	Y	S	T	−0.89
C	Y	Q	S	⋈	Y	S	T	M	−0.79
Y	Q	S	Y	⋈	S	T	M	S	−0.52
Q	S	Y	S	⋈	T	M	S	I	−0.75
S	Y	S	T	⋈	M	S	I	T	−0.54
Y	S	T	M	⋈	S	I	T	D	−0.69
S	T	M	S	⋈	I	T	D	C	−0.60
T	M	S	I	⋈	T	D	C	R	−0.98
M	S	I	T	⋈	D	C	R	E	−0.99
S	I	T	D	⋈	C	R	E	T	−0.42
I	T	D	C	⋈	R	E	T	G	−0.51
T	D	C	R	⋈	E	T	G	S	−0.93
D	C	R	E	⋈	T	G	S	S	−0.59
C	R	E	T	⋈	G	S	S	K	−0.90
R	E	T	G	⋈	S	S	K	Y	−0.52
E	T	G	S	⋈	S	K	Y	P	−3.71
T	G	S	S	⋈	K	Y	P	N	−0.90
G	S	S	K	⋈	Y	P	N	C	−0.74
S	S	K	Y	⋈	P	N	C	A	−3.48
S	K	Y	P	⋈	N	C	A	Y	−0.60
K	Y	P	N	⋈	C	A	Y	K	−0.86
Y	P	N	C	⋈	A	Y	K	T	−0.85
P	N	C	A	⋈	Y	K	T	T	−3.54
N	C	A	Y	⋈	K	T	T	Q	−0.71
C	A	Y	K	⋈	T	T	Q	A	−1.08
A	Y	K	T	⋈	T	Q	A	N	−3.81
Y	K	T	T	⋈	Q	A	N	K	−0.88
K	T	T	Q	⋈	A	N	K	H	−0.58
T	T	Q	A	⋈	N	K	H	I	−3.96
T	Q	A	N	⋈	K	H	I	I	−0.09
Q	A	N	K	⋈	H	I	I	V	−0.42
A	N	K	H	⋈	I	I	V	A	−3.54
N	K	H	I	⋈	I	V	A	C	−0.49
K	H	I	I	⋈	V	A	C	E	−1.07
H	I	I	V	⋈	A	C	E	G	−1.20
I	I	V	A	⋈	C	E	G	N	−0.62
I	V	A	C	⋈	E	G	N	P	−1.14
V	A	C	E	⋈	G	N	P	U	−1.06
A	C	E	G	⋈	N	P	Y	V	−0.87
C	E	G	N	⋈	P	Y	V	P	−0.70

(*continued*)

TABLE II. The 239 Noncleavable Peptides by HIV-1 Protease* (*Continued*)

Peptide sequence†									
R_4	R_3	R_2	R_1	⋈	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	$\Delta^\ddagger$
E	G	N	P	⋈	Y	V	P	V	−0.05
G	N	P	Y	⋈	V	P	V	H	−1.01
N	P	Y	V	⋈	P	V	H	F	−0.88
P	Y	V	P	⋈	V	H	F	D	−0.77
Y	V	P	V	⋈	H	F	D	A	−1.17
V	P	V	H	⋈	F	D	A	S	−0.21
P	V	H	F	⋈	D	A	S	V	−0.90

*Of them 122 octapeptides are extracted from hen egg lysozyme and 117 from bovine pancreatic ribonuclease since neither of the two proteins have showed any cleavage sites even if they are completely denatured to make any part of them is accessible to the active site of HIV protease.

†The symbol ⋈ means the peptide bond between amino acids at R_1 and $R_{1'}$ is noncleavable by HIV-1 protease. ‡See footnote † to Table I.

within a limited region is highly favored over the others, cleavage at this site will result in fragments that are too small to serve as substrates, thereby removing the other predicted cleavages from the picture. Experimentally, we have been able to document mixtures of peptides resulting in such a case where the propensity for cleavage at one or the other site was about the same.[26–28] It is not clear how much more favorable a cleavage point needs to be in order to prevent experimental observation of hydrolysis at nearby susceptible sites.

With this preamble in mind, the new algorithm correctly predicts all 8 of the processing sites hydrolyzed by the HIV-1 protease in the HIV-1 gag and gag/pol polyproteins during the course of viral maturation. Of course, this follows from the fact that these sequences are part of the testing series given in Table V, all of which are predicted. In order to test overprediction by the method, the whole *pol* gene product was analyzed to see how many sites of hydrolysis would be predicted. In this case, 210 cleavage sites out of a possible 988 sites were predicted based upon the criterion that + values of Δ are predictive for cleavage, and − values are nonpredictive! At first sight, this would suggest that the algorithm is overpredictive to a degree so as to make it impractical. However, it is clear that the majority of these predicted sites must be in adjacent regions and based upon arguments presented above, a high probability cleavage will dominate so as to remove sites of lower probability from experimental consideration. Therefore, we sought to define a cutoff value of somewhere in the range of 0.5 to 1.0 so as to allow prediction of a fewer number of higher probability sites. Using a cutoff value whereby Δ values > 0.8 are predictive for cleavage, one obtains 46 predicted sites of hydrolysis. These sites are presented in Table VI. These sites will be discussed, in turn, relative to their occurrence in specific proteins encoded within the pol gene.

Earlier studies from our laboratory have documented sites of cleavage by the HIV-1 protease in partially unfolded reverse transcriptase (RT)[2] and protease.[18] The sites in the protease are, in fact, observed naturally as sites of autolysis.[18,29–31] In this case, the algorithm predicts three sites of hydrolysis that were not seen, one involving a Thr-Leu bond directly adjacent to an observed site of hydrolysis. Here, the value for the Thr-Leu bond is significantly lower than that for the Leu–Trp (Table VI). The Δ value for the predicted but not cleaved Gly–Gly bond is high (1.48), but, in fact, such a cleavage has never been observed. The physical state of the protease as its own substrate is not clearly understood, and it may be that these refractory bonds reside in parts of the autolyzed monomer that retain enough folded structure to limit accessibility.

In the case of RT/RNase H, 8 of the 23 sites predicted were observed experimentally as hydrolysis sites in the partially denatured molecule.[2] Interestingly, most of the observed cleavage sites have high Δ values, ranging from 1.79 to 3.55. Overprediction could result from a number of factors as discussed above. Bond cleavage was seen only when the pH of the RT solution was lowered to 4, a rather mild condition of denaturation which could easily sustain folding in substantial portions of the molecule. Therefore, site predicted, but not seen could have been shielded from access by the enzyme. In fact, many of these sites are located in large segments of the RT hydrolysis products which might be expected to retain folded structure. This might apply to the region from 390 to 458, where 7 sites are predicted but not hydrolyzed. Of course the Glu-457–Leu bond with $\Delta = 0.96$ loses out to the adjacent Leu-458–Ala bond ($\Delta = 2.21$) which is a prominent site of hydrolysis. In the RNase H region all but one of the predicted sites were observed to be hydrolyzed. The Glu-701–Gln bond predicted but not observed has a Δ value of 1.22, but again, this bond has never been seen to be cleaved by the HIV-1 protease.

Seven sites in the integrase are predicted in addi-

TABLE III. Probability of Finding an Amino Acid at Each of the Eight Subsites for the 62 Cleavable Peptides by HIV-1 Protease*

Amino acid	R_4	R_3	R_2	R_1	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$
X	$P_4^+(X)$	$P_3^+(X)$	$P_2^+(X)$	$P_1^+(X)$	$P_{1'}^+(X)$	$P_{2'}^+(X)$	$P_{3'}^+(X)$	$P_{4'}^+(X)$
A	0.15	0.05	0.13	0.06	0.18	0.08	0.11	0.03
C	0.00	0.00	0.02	0.02	0.00	0.00	0.02	0.02
D	0.08	0.05	0.03	0.02	0.00	0.00	0.05	0.10
E	0.06	0.18	0.06	0.02	0.06	0.47	0.13	0.03
F	0.02	0.06	0.00	0.27	0.11	0.00	0.08	0.06
G	0.15	0.05	0.02	0.06	0.03	0.00	0.03	0.13
H	0.03	0.00	0.02	0.00	0.00	0.00	0.05	0.03
I	0.00	0.03	0.19	0.00	0.05	0.05	0.05	0.02
K	0.03	0.05	0.00	0.00	0.00	0.00	0.00	0.08
L	0.05	0.13	0.06	0.29	0.08	0.08	0.03	0.06
M	0.00	0.02	0.02	0.06	0.03	0.00	0.05	0.03
N	0.00	0.00	0.15	0.08	0.00	0.00	0.03	0.03
P	0.10	0.02	0.00	0.00	0.16	0.00	0.00	0.08
Q	0.05	0.13	0.00	0.00	0.00	0.13	0.03	0.06
R	0.06	0.08	0.00	0.00	0.05	0.00	0.06	0.05
S	0.11	0.03	0.03	0.00	0.02	0.02	0.06	0.10
T	0.05	0.06	0.08	0.00	0.05	0.05	0.10	0.05
V	0.03	0.00	0.19	0.00	0.05	0.13	0.06	0.02
W	0.00	0.05	0.00	0.00	0.03	0.00	0.02	0.00
Y	0.02	0.02	0.00	0.11	0.10	0.00	0.03	0.00

*The sequences of these peptides are given in Table I.

TABLE IV. Probability of Finding an Amino Acid at Each of the Eight Subsites for the 239 Noncleavable Peptides by HIV-1 Protease.*

Amino acid	R_4	R_3	R_2	R_1	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$
X	$P_4^-(X)$	$P_3^-(X)$	$P_2^-(X)$	$P_1^-(X)$	$P_{1'}^-(X)$	$P_{2'}^-(X)$	$P_{3'}^-(X)$	$P_{4'}^-(X)$
A	0.10	0.10	0.10	0.10	0.09	0.09	0.09	0.09
C	0.06	0.06	0.06	0.06	0.06	0.07	0.06	0.06
D	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05
E	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.02
F	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
G	0.06	0.06	0.06	0.06	0.06	0.06	0.06	0.06
H	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.02
I	0.03	0.03	0.04	0.04	0.04	0.04	0.04	0.04
K	0.07	0.06	0.06	0.06	0.06	0.06	0.06	0.05
L	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
M	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
N	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
P	0.03	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Q	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04
R	0.05	0.05	0.05	0.06	0.06	0.05	0.06	0.06
S	0.10	0.10	0.10	0.10	0.10	0.10	0.10	0.10
T	0.07	0.07	0.07	0.07	0.07	0.07	0.07	0.07
V	0.05	0.06	0.05	0.05	0.05	0.05	0.05	0.06
W	0.02	0.03	0.03	0.03	0.03	0.03	0.03	0.03
Y	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04

*The sequences of these peptides are given in Table II.

tion to the natural processing site at residue 715 (Table VI), but we have no experimental data on the course of hydrolysis of the denatured integrase by the HIV-1 protease to either refute or substantiate these predictions.

Finally, in a recent study by Fan et al.[28] some interesting observations were made relative to the way the HIV-1 protease hydrolyzes the HIV-1 and HIV-2 RTs. Given the two homologous stretches of sequence:

TABLE V. The Predicted Results for the 55 Testing Peptides Known Cleavable by HIV-1 Protease.*

Peptide sequence and cleavage site										
R_4	R_3	R_2	R_1	⇓	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	Δ[†]	$h - 0.13$[‡]
T	Q	N	Y	⇓	P	I	V	Q[§]	1.76	0.09
S	N	N	Y	⇓	P	I	V	Q	1.04	−0.11
S	K	N	Y	⇓	P	I	V	Q	1.08	−0.07
S	Q	N	F	⇓	P	I	V	Q	1.59	0.55
S	Q	N	Y	⇓	L	I	V	Q	1.12	0.04
S	Q	N	Y	⇓	T	I	V	Q	0.83	0.03
S	Q	N	Y	⇓	P	I	I	Q	0.78	0.43
S	Q	N	Y	⇓	P	I	E	Q	0.90	0.50
S	Q	N	Y	⇓	P	I	V	P	1.18	0.55
S	Q	N	Y	⇓	P	I	V	E	1.01	0.14
T	F	N	F	⇓	P	Q	I	T	1.00	0.71
Y	F	N	F	⇓	P	Q	I	T	1.06	0.69
S	C	N	F	⇓	P	Q	I	T	0.85	0.62
S	Y	N	F	⇓	P	Q	I	T	1.00	0.40
S	F	T	F	⇓	P	Q	I	T	1.47	0.72
S	F	Y	F	⇓	P	Q	I	T	0.93	0.26
S	F	N	S	⇓	P	Q	I	T	0.64	−0.01
S	F	N	Y	⇓	P	Q	I	T	1.14	0.64
S	F	N	F	⇓	G	Q	I	T	0.82	0.44
S	F	N	F	⇓	L	Q	I	T	1.03	0.66
S	F	N	F	⇓	P	P	I	T	1.16	0.16
S	F	N	F	⇓	P	L	I	T	1.26	0.65
S	F	N	F	⇓	P	Q	V	T	1.62	0.72
S	F	N	F	⇓	P	Q	D	T	1.25	0.74
S	F	N	F	⇓	P	Q	I	I	1.16	0.38
S	Q	N	Y	⇓	P	A	V	Q**	1.00	0.07
S	Q	N	Y	⇓	P	N	V	Q	0.83	−0.07
S	Q	N	Y	⇓	P	L	V	Q	1.14	0.21
S	Q	N	Y	⇓	P	V	V	Q	1.01	0.25
S	Q	N	Y	⇓	P	I	L	Q[††]	1.17	−0.06
S	Q	N	Y	⇓	P	I	F	Q	1.23	0.56
S	Q	L	Y	⇓	P	I	V	Q	1.34	0.01
S	Q	C	Y	⇓	P	I	V	Q	1.02	0.09
S	Q	A	Y	⇓	P	I	V	Q	0.89	0.04
S	Q	T	Y	⇓	P	I	V	Q	1.11	0.11
S	Q	N	M	⇓	P	I	V	Q	0.90	0.36
A	R	V	L	⇓	F	E	A	L[‡‡]	1.13	0.72
A	R	V	L	⇓	F	Q	A	L	0.67	0.61
A	R	V	L	⇓	F	I	A	L	1.07	0.39
A	R	V	L	⇓	F	V	A	L	0.89	0.39
A	R	V	L	⇓	F	A	A	L	1.03	0.17
A	R	V	L	⇓	F	D	A	L	0.79	−0.05
A	R	V	L	⇓	F	N	A	L	0.83	−0.04
A	R	V	L	⇓	F	T	A	L	1.01	0.11
A	R	N	L	⇓	F	E	A	L	0.82	0.73
A	R	N	L	⇓	F	Q	A	L	0.36	0.62
A	R	N	L	⇓	F	I	A	L	0.76	0.40
A	R	N	L	⇓	F	V	A	L	0.58	0.40
A	R	V	Y	⇓	P	E	A	L	1.47	0.63
A	R	N	Y	⇓	P	E	A	L	1.13	0.64
S	Q	N	Y	⇓	P	I	V		1.01	0.36
S	Q	N	Y	⇓	P	I	V	L	1.01	0.34
A	R	N	Y	⇓	P	I	V	L	0.93	0.14
A	Q	N	Y	⇓	P	I	V	L	1.05	0.35
R	Q	N	Y	⇓	P	I	A	L	1.45	0.41

*These peptides are not included in the training database of Table I and hence form an independent testing set for the enzyme.

[†] See footnote † to Table I.

[‡] See footnote ‡ to Table I.

[§] The following 25 entries are from Partin et al.[23]

**The following 4 entries are from Bláha et al.[24]

[††] The following 7 entries are from Tözsér et al.[9]

[‡‡] The following 19 entries are from Griffiths et al.[8]

TABLE VI. Predicted HIV-1 Cleavage Sites in POL_HV1H2.SW[*]

Cleavage site	Residue[†] No.	Δ score	Experimentally observed	*pol* protein
FRED-LAFL	5	1.10		
REDL-AFLQ	6	1.13		
SPSE-AGAD	43	1.11		
SFNF ‖ PQVT	56	1.62	+	Protease
PQVT-LWQR	60	0.82		
QVTL-WQRP	61	1.26	+	
DTVL-EEMS	89	2.03	+	
KMIG-GIGG	104	1.48		
DQIL-IEIC	119	1.45	+	
QIGC-TLNF	151	1.20		
TLNF ‖ PISP	155	2.17	+	RT
QWPL-TEEK	181	1.27		
WPLT-EEKI	182	0.80		
IKAL-VEIC	189	1.17		
VEIC-TEME	193	3.55	+	
TQDF-WEVQ	242	2.09	+	
DVGD-AYFS	268	1.21		
GDAY-FSVP	270	2.47	+	
RQHL-LRWG	364	1.20		
KEPP-FLWM	381	0.83		
LWMG-YELH	386	2.56	+	
YELH-PDKW	390	0.93		
PIVL-PEKD	401	0.82		
LNWA-SQIY	422	0.87		
TKAL-TEVI	444	0.98		
VIPL-TEEA	450	1.11		
LTEE-AELE	453	1.49		
AELE-LAEN	457	0.96		
ELEL/AENR	458	2.21	+	
SKDL-IAEI	480	1.86	+	
VKQL-TEAV	523	1.05		
ETWW-TEYW	557	0.83		
PIVG-AETF	591	1.79	+	
AETF ‖ YVDG	595	1.98	+	RNase H
QAIY-LALQ	638	1.70	+	
LEVN-IVTD	649	0.93	+	
EKVY-LAWV	687	1.98	+	
GGNE-QVDK	701	1.22		
RKVL ‖ FLDG	715	1.89	+	Integrase
LKGE-AMHG	763	1.69	?	
QLDC-THLE	780	0.80	?	
CTHL-EGKV	783	0.83	?	
GYIE-AEVI	800	1.78	?	
YNPQ-SQGV	861	0.80	?	
IQNF-RVYY	948	0.95	?	
FRVY-YRDS	941	1.17	?	

[*] Amino acid sequence (accession number P04585) extracted from SWISS-PROT database available through the Wisconsin Package. Genetics Program Group, *Program Manual for the GCG Package,* Version 7, (April, 1991), 575 Science Drive, Madison, WI, USA 53711. ‖, known N-terminus of the mature protein named in the right-hand column; ?, predicted cleavable sites in integrase, but not yet tested experimentally.

[†] Number denotes R_1 amino acid.

HIV-1 RT: ...Gln–**Ala**$_{481}$–Ile–Tyr–⇓–
Leu–?–**Ala–Leu**–Gln–**Asp...**
HIV-2 RT: ...Glu–**Ala**$_{481}$–Phe–Ala–?–Met–⇓–
Ala–Leu–Thr–**Asp...** (7)

the protease selects the bonds indicated by the arrows. The algorithm correctly predicts these bonds as cleavable, with Δ values of 1.70 and 0.68, respectively. Would the algorithm also predict cleavage of the Leu–Ala bond adjacent to the cleavable Tyr–Leu bond in HIV-1 RT and the Ala–Met bond adjacent to the cleavable Met–Ala bond in HIV-2 RT [see Eq. (7)]? The Δ value for both of them is 0.04, very close to zero, and hence their priority of being selected for cleavage by the enzyme would be too low to be observed, especially in competition with the adjacent highly cleavable bonds.

Another interesting difference displayed by the HIV-1 protease in its cleavage preference[28] is seen in the following pair:

$$\begin{cases}\text{HIV-1 RT:} & \ldots\textbf{Gln–Tyr}\text{–Ala–Leu}_{503}\bowtie \\ & \textbf{Gly–Ile}\text{–Ile–Gln}\ldots \\ \text{HIV-2 RT:} & \ldots\textbf{Gln–Tyr}\text{–Val–Met}_{503}\text{–}\Downarrow\text{–} \\ & \textbf{Gly–Ile}\text{–Val–Ala}\ldots\end{cases} \qquad (8)$$

Here, as indicated by the arrow, the Met–Gly bond HIV-2 RT is hydrolyzed, whereas the corresponding Leu–Ala bond in HIV-1 RT is not. In this case, however, neither bond, Met–Gly or Leu–Gly, is predicted as a site of cleavage.

CONCLUSION

Understanding the specificity of the HIV protease is basic to development of inhibitors of the enzyme, and the attempt to define protease inhibitors represents a considerable effort in the search for drugs against AIDS. The present algorithm constitutes an improvement over earlier predictive schemes in providing a closer agreement between predicted and experimentally observed sites of hydrolysis by the protease. It, therefore, expands our understanding of structural determinants that are of paramount importance in defining substrates of this important enzyme.

ACKNOWLEDGMENTS

Valuable discussions with Dr. Ferenc J. Kézdy and Professor Chun-Ting Zhang are gratefully acknowledged. The authors also thank Joan V. Baker for helping draw Figure 1.

REFERENCES

1. Henderson, L.E., Benveniste, R.E., Sowder, R.C., Copeland, T.D., Schutz, A.M., Oroszlan, S. Molecular characterization of *gag* proteins from simian immunodeficiency virus (SIV$_{\text{Mne}}$). J. Virol. 62:2587–2595, 1988.
2. Tomasselli, A.G., Sarcich, J.L., Barrett, L.J., Reardon, I.M., Howe, W.J., Evans, D.B., Sharma, S.K., Heinrikson, R.L. Human immunodeficiency virus type-1 reverse transcriptase and ribonuclease H as substrates of the viral protease. Protein Sci. 2:2167–2176, 1993.
3. Tomasselli, A.G., Heinrikson, R.L. Specificity of retroviral proteases: An analysis of viral and nonviral protein substrates. Methods Enzymol. 241:279–301, 1994.
4. Dunn, B.M., Gustchina, A., Wlodaver, A., Kay, J. Subsite preferences of retroviral proteinases. Methods Enzymol. 241:254–278, 1994.
5. Konvalinka, J., Strop, P., Velek, J., Cerna, V., Kostka, V., Phylip, L.H., Richards, A.D., Dunn, B.M., Kay, J. Sub-site preferences of the aspartic proteinase from the HIV-1 virus. FEBS Lett. 268:35–38, 1990.
6. Margolin, N., Heath, W., Osborne, E., Lai, M., Vlahos, C. Substitution at the $P_{2'}$ site of gag p17–p24 affect cleavage efficiency by HIV-1 protease. Biochem. Biophys. Res. Commun. 167:554–560, 1990.
7. Phylip, L.H., Richards, A.D., Konvalinka, J.K., Strop, P., Blaha, I., Velek, J., Kostka, V., Ritchie, A.J., Broadhurst, A.V., Farmerie, W.J., Scarborough, P.E., Dunn, B.M. Hydrolysis of synthetic chromogenic substrates by HIV-1 and HIV-2 proteinases. Biochem. Biophys. Res. Commun. 171: 439–444, 1990.
8. Griffiths, J.T., Phylip, L.H., Konvalinka, J., Strop, P., Gustchina, A., Wlodawer, A., Davenport, R., Briggs, R., Dunn, B.M., Kay, J. Different requirements for productive interaction between the active site of HIV-1 protease and substrates containing -hydrophobic*hydrophobic- or -aromatic*pro- cleavage sites. Biochemistry 31:5193–5200, 1992.
9. Tözsér, J., Weber, I.T., Gustchina, A., Bláha, I., Copeland, T.D., Louis, J.M., Oroszlan, S. Kinetic and modeling studies of S_3–S_3' studies of HIV proteinases. Biochemistry 31: 4793–4800, 1992.
10. Poorman, R.A., Tomasselli, A.G., Heinrikson, R.L., Kézdy, F.J. 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. 266:14554–14561, 1991.
11. Chou, K.C., Zhang, C.T., Kézdy, F.J. A vector projection approach to predicting HIV protease cleavage sites in proteins. Proteins 16:195–204, 1993.
12. Chou, J.J. Predicting cleavability of peptide sequences by HIV protease via correlation-angle approach. J. Protein Chem. 12:291–302, 1993.
13. Chou, K.C., Zhang, C.T. Studies on the specificity of HIV protease: An application of Markov chain theory. J. Protein Chem. 12:709–724, 1993.
14. Zhang, C.T., Chou, K.C. An alternate-subsite-coupled model for predicting HIV protease sites in proteins. Protein Eng. 7:65–73, 1994.
15. Chou, K.C. A vectorized sequence-coupling model for predicting HIV protease cleavage sites in proteins. J. Biol. Chem. 268:16938–16948, 1993.
16. Schechter, I., Berger, A. On the size of the active site in proteases. I. Papain. Biochem. Biophys. Res. Commun. 27: 157–162, 1967.
17. Chou, K.C., Zhang, C.T. A correlation-coefficient method to predicting protein-structural classes from amino acid compositions. Eur. J. Biochem. 207:429–433, 1992.
18. Mildner, A.M., Rothrock, D.J., Leone, J., Bannow, C.A., Lull, J.M., Reardon, I.M., Sarcich, J.L., Howe, W.J., Tomich, C-S.C., Smith, C.W., Heinrikson, R.L., Tomasselli, A.G. The HIV protease as enzyme and substrate: Mutagenesis of autolysis sites and generation of a stable mutant with retained kinetic properties. Biochemistry 33:9405–9413, 1994.
19. Riviere, Y., Blank, V., Kourilsky, P., Israel, A. Processing of the precurson of NF-κB by HIV-1 protease during acute infection. Nature (London) 350:625–626, 1991.
20. Tomaszek, T.A., Jr., Moore, M.L., Strickler, J.E., Sanchez, R.I., Dixon, J.S., Metcalf, B.W., Hassell, A., Dreyer, G.B., Brooks, I., Debouck, C., Meek, T.D. Proteolysis of an active site peptide of lactate dehydrogenase by human immunodeficiency virus type 1 protease. Biochemistry 31:10153–10168, 1992.
21. Chattopadhyay, D., Evans, D.B., Deibel, M.R., Jr., Vosters, A.F., Eckenrode, F.M., Einspair, H.M., Hui, J.O., Tomasselli, A.G., Zurcher-Neely, H.A., Heinrikson, R.L., Sharma, S.K. Purification and characterization of heterodimeric HIV-1 reverse transcriptase produced by in vivo processing of p66 with recombinant HIV-1 protease. J. Biol. Chem. 267:14227–14232, 1992.
22. Oswald, von der Helm Fibronectin is a non-viral substrate for the HIV protease. FEBS Lett. 292:298–300, 1991.
23. Partin, K., Kräusslich, H.G., Ehrlich, L., Wimmer, E., Carter, C. Mutational analysis of a native substrate of the human immunodeficiency virus type 1 proteinase. J. Virol 64:3938–3947, 1990.
24. Bláha, I., Nemec, J., Tözsér, J., Oroszlan, S. Synthesis of homologous peptides using fragment condensation: Analogs of an HIV proteinase substrate. Int. J. Peptide Protein Res. 38:453–458, 1992.
25. Darke, P.L., Nutt, R.F., Brady, S.F., Garsky, V.M., Ciccarone, T.M., Leu, C-T., Lumma, P.K., Freidinger, R.M., Veber, D.F., Sigal, I.S. HIV-1 protease specificity of peptide cleavage is sufficient for processing of GAG and POL polyproteins. Biochem. Biophys. Res. Commun. 156:297–303, 1988.
26. Tomasselli, A.G., Howe, W.J., Hui, J.O., Sawyer, T.K., Reardon, I.M., DeCamp, D.L., Craik, C.S., Heinrikson,

R.L. Calcium-free calmodulin is a substrate of proteases from human immunodeficiency viruses 1 and 2. Proteins 10:1–9, 1991.

27. Tomasselli, A.G., Hui, J.O., Adams, L., Chosay, J., Lowery, D., Greenberg, B., Yem, A., Deibel, M.R., Zurcher-Neely, H., Heinrikson, R.L. Actin, troponin C, Alzheimer amyloid precursor protein and pro-interleukin 1β as substrates of the HIV-1 protease. J. Biol. Chem. 266:14548–14553, 1991.

28. Fan, N., Rank, K.B., Leone, J.W., Heinrikson, R.L., Bannow, C.A., Smith, C.W., Evans, D.B., Poppe, S.M., Tarpley, W.G., Rothrock, D.J., Tomasselli, A.G., Sharma, S.K. The differential processing of homodimers of reverse transcriptase from human immunodeficiency virus type 1 and 2 is a consequence of the distinct specificities of the viral proteases. J. Biol. Chem., 270:13573–13579, 1995.

29. Strickler, J.E., Gorniak, J., Dayton, B., Meek, T., Moore, M., Magaard, V., Malinowski, Debouck, C. Characterization and autoprocessing of precursor and mature forms of human immunodeficiency virus type 1 (HIV 1) protease purified from Escherichia coli. Proteins 6:139–154, 1989.

30. Hui, J.O., Tomasselli, A.G., Reardon, I.M., Lull, J.M., Brunner, D.P., Tomich, C-s C., Heinrikson, R.L. Large scale purification and refolding of HIV-1 protease from Escherichia coli inclusion bodies. J. Prot. Chem. 12:323–327, 1993.

31. Rosé, J.R., Salto, R., Craik, C.S. Regulation of autoproteolysis of the HIV-1 and HIV-2 proteases with engeeniered amino acid substitutions. J. Biol. Chem. 268:11939–11945, 1993.

32. Bhat, U.N. Elements of Applied Stochastic processes. Chapt. 3. New York: Wiley, 1984.

Appendix A
The Conditional Probabilities Derived From the 62 Cleavable Peptides by HIV-1 Protease

1. For the subsite R_3: $P_3^+(X_3\|X_4)$, where X_3 represents the amino acids along the row, and X_4 the amino acid along the column																				
	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.00	0.00	0.00	0.56	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.22	0.00	0.11	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D	0.20	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.20	0.20	0.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00
F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00
G	0.00	0.00	0.22	0.00	0.00	0.22	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.11	0.00	0.00	0.22	0.00
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
K	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.33	0.00
M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P	0.00	0.00	0.00	0.00	0.33	0.17	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.17	0.00	0.00	0.00
Q	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.00	0.14	0.29	0.00	0.00	0.00	0.14	0.14	0.00	0.00	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00
V	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
2. For the subsite R_2: $P_2^+(X_2\|X_3)$, where X_2 represents the amino acids along the row, and X_3 the amino acid along the column																				
A	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.18	0.09	0.00	0.18	0.00	0.00	0.00	0.09	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00	0.18	0.18	0.00	0.00
F	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.50	0.00	0.00
K	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00
L	0.13	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.38	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00
M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Q	0.13	0.00	0.13	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.13	0.25	0.00	0.00
R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.20	0.00	0.00	0.00	0.40	0.00	0.20	0.00	0.00
S	0.50	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.25	0.00	0.00
V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3. For the subsite R_1: $P_1^+(X_1\|X_2)$, where X_1 represents the amino acids along the row, and X_2 the amino acid along the column																				
A	0.25	0.00	0.13	0.00	0.13	0.00	0.00	0.00	0.00	0.13	0.13	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
C	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.08	0.00	0.00	0.17	0.08	0.00	0.00	0.00	0.33	0.17	0.08	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08
K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.25	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25
M	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.00	0.00	0.00	0.00	0.44	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.22
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.20	0.00	0.00	0.00	0.60	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V	0.08	0.00	0.00	0.08	0.08	0.08	0.00	0.00	0.00	0.25	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(continued)

Appendix A (*Continued*)

The Conditional Probabilities Derived From the 62 Cleavable Peptides by HIV-1 Protease

4. For the subsite $R_{1'}$: $P_{1'}^{+}(X_{1'}|X_1)$, where $X_{1'}$ represents the amino acids along the row, and X_1 the amino acid along the column

A	0.25	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
D	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F	0.00	0.00	0.00	0.12	0.12	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.29	0.00	0.12	0.00	0.00	0.06	0.06	0.18
G	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.25
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.28	0.00	0.00	0.06	0.11	0.06	0.00	0.11	0.00	0.06	0.00	0.00	0.06	0.00	0.06	0.00	0.06	0.06	0.06	0.06
M	0.25	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00
N	0.00	0.00	0.00	0.00	0.20	0.20	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.14	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.29	0.00	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.00	0.00

5. For the subsite $R_{2'}$: $P_{2'}^{+}(X_{2'}|X_{1'})$, where $X_{2'}$ represents the amino acids along the row, and $X_{1'}$ the amino acid along the column

A	0.00	0.00	0.00	0.73	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.09	0.00	0.00	0.09	0.00	0.00	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F	0.14	0.00	0.00	0.29	0.00	0.00	0.00	0.00	0.00	0.43	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00
G	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.33	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00
K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.40	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.20	0.00	0.00
M	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P	0.00	0.00	0.00	0.10	0.00	0.00	0.00	0.20	0.00	0.10	0.00	0.00	0.00	0.20	0.00	0.00	0.10	0.30	0.00	0.00
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
R	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00
S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
T	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
V	0.00	0.00	0.00	0.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00
W	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.17	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.33	0.00	0.00

6. For the subsite $R_{3'}$: $P_{3'}^{+}(X_{3'}|X_{2'})$, where $X_{3'}$ represents the amino acids along the row, and $X_{2'}$ the amino acid along the column

A	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20	0.00
C	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.17	0.00	0.00	0.17	0.07	0.03	0.10	0.07	0.00	0.03	0.07	0.03	0.00	0.00	0.03	0.03	0.14	0.03	0.00	0.00
F	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.33	0.00	0.00
K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.00	0.00	0.20	0.20	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20
M	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	0.00	0.00	0.00	0.00	0.13	0.00	0.00	0.13	0.00	0.00	0.13	0.00	0.00	0.13	0.25	0.13	0.00	0.13	0.00	0.00
R	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
S	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
T	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.33
V	0.25	0.13	0.13	0.00	0.13	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.13	0.00	0.00	0.13	0.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Y	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

(continued)

Appendix A (*Continued*)

The Conditional Probabilities Derived From the 62 Cleavable Peptides by HIV-1 Protease

7. For the subsite $R_{4'}$: $P_{4'}^{+}(X_{4'}|X_{3'})$, where $X_{4'}$ represents the amino acids along the row, and $X_{3'}$ the amino acid along the column

A	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.14	0.29	0.14	0.00	0.14	0.14	0.00	0.00	0.00	0.00	0.00	0.00
C	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
D	0.00	0.00	0.00	0.00	0.00	0.67	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
E	0.14	0.00	0.14	0.14	0.14	0.00	0.00	0.14	0.14	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.00	0.00	0.00
F	0.00	0.00	0.40	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.40	0.00	0.00	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
H	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.33	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00
K	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
L	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00
M	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00	0.33	0.00	0.00	0.00	0.00
N	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00
P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Q	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00
R	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.25	0.25	0.00	0.00	0.25	0.00	0.00	0.00	0.00
S	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.00	0.25	0.25	0.00	0.00	0.00	0.00
T	0.00	0.00	0.17	0.00	0.33	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.17	0.17	0.00	0.00	0.00
V	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.25	0.50	0.00	0.00	0.25	0.00	0.00	0.00
W	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
Y	0.00	0.00	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.50	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Appendix B
The conditional Probabilities Derived From the 239 Noncleavable Peptides by HIV-1 Protease

1. For the subsite R_3: $P_3^-(X_3|X_4)$, where X_3 represents the amino acids along the row, and X_4 the amino acid along the column

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.26	0.09	0.04	0.00	0.00	0.00	0.00	0.00	0.13	0.04	0.04	0.04	0.00	0.00	0.00	0.09	0.04	0.04	0.13	0.04
C	0.20	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.07	0.13	0.00	0.00	0.00	0.07
D	0.00	0.09	0.00	0.00	0.00	0.27	0.00	0.09	0.00	0.00	0.00	0.09	0.00	0.00	0.09	0.09	0.00	0.18	0.00	0.09
E	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.14	0.29	0.29	0.00	0.00	0.00
F	0.00	0.00	0.00	0.40	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.07	0.07	0.21	0.00	0.07	0.14	0.21	0.07	0.00	0.00	0.07
H	0.00	0.00	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.13	0.00	0.13	0.13	0.00	0.00	0.00	0.25	0.25	0.00	0.00
K	0.00	0.00	0.06	0.06	0.13	0.06	0.06	0.06	0.06	0.00	0.00	0.13	0.06	0.00	0.06	0.06	0.06	0.06	0.00	0.06
L	0.22	0.11	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.00	0.00	0.00
M	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
N	0.04	0.13	0.04	0.00	0.04	0.08	0.00	0.04	0.04	0.08	0.00	0.00	0.04	0.04	0.08	0.04	0.13	0.04	0.04	0.08
P	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17
Q	0.40	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.10	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00
R	0.00	0.23	0.00	0.08	0.00	0.08	0.08	0.00	0.00	0.00	0.00	0.31	0.00	0.08	0.00	0.00	0.08	0.00	0.08	0.00
S	0.08	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.04	0.08	0.00	0.08	0.00	0.04	0.13	0.21	0.13	0.04	0.00	0.04
T	0.12	0.00	0.24	0.00	0.06	0.06	0.00	0.00	0.06	0.00	0.06	0.12	0.06	0.12	0.00	0.06	0.06	0.00	0.00	0.00
V	0.23	0.15	0.00	0.00	0.08	0.00	0.08	0.00	0.00	0.00	0.00	0.15	0.08	0.15	0.00	0.08	0.00	0.00	0.00	0.00
W	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.40	0.20	0.00
Y	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.11	0.11	0.22	0.00	0.11	0.00	0.00

2. For the subsite R_2: $P_2^-(X_2|X_3)$, where X_2 represents the amino acids along the row, and X_3 the amino acid along the column

A	0.26	0.09	0.04	0.00	0.00	0.00	0.00	0.00	0.13	0.04	0.04	0.04	0.00	0.00	0.00	0.09	0.04	0.04	0.13	0.04
C	0.20	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.07	0.13	0.00	0.00	0.00	0.07
D	0.00	0.09	0.00	0.00	0.00	0.27	0.00	0.09	0.00	0.00	0.00	0.09	0.00	0.00	0.09	0.09	0.00	0.18	0.00	0.09
E	0.00	0.00	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.14	0.00	0.00	0.00	0.00	0.14	0.29	0.29	0.00	0.00	0.00
F	0.00	0.00	0.00	0.40	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.07	0.07	0.21	0.00	0.07	0.14	0.21	0.07	0.00	0.00	0.07
H	0.00	0.00	0.00	0.25	0.00	0.25	0.00	0.25	0.00	0.00	0.25	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.00	0.13	0.00	0.13	0.13	0.00	0.00	0.00	0.25	0.25	0.00	0.00
K	0.00	0.00	0.07	0.00	0.14	0.07	0.07	0.07	0.07	0.00	0.00	0.14	0.07	0.00	0.07	0.07	0.07	0.00	0.00	0.07
L	0.22	0.11	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.00	0.00	0.00
M	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
N	0.04	0.13	0.04	0.00	0.04	0.08	0.00	0.04	0.04	0.08	0.00	0.00	0.04	0.04	0.08	0.04	0.13	0.04	0.04	0.08
P	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17
Q	0.40	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.10	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00
R	0.00	0.23	0.00	0.08	0.00	0.08	0.08	0.00	0.00	0.00	0.00	0.31	0.00	0.08	0.00	0.00	0.08	0.00	0.08	0.00
S	0.08	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.04	0.08	0.00	0.08	0.00	0.04	0.13	0.21	0.13	0.04	0.00	0.04
T	0.12	0.00	0.24	0.00	0.06	0.06	0.00	0.00	0.06	0.00	0.00	0.06	0.12	0.06	0.12	0.00	0.06	0.06	0.00	0.00
V	0.21	0.14	0.00	0.00	0.07	0.00	0.14	0.00	0.00	0.00	0.00	0.14	0.07	0.14	0.00	0.07	0.00	0.00	0.00	0.00
W	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.33	0.17	0.00
Y	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.11	0.11	0.22	0.00	0.11	0.00	0.00

3. For the subsite R_1: $P_1^-(X_1|X_2)$, where X_1 represents the amino acids along the row, and X_2 the amino acid along the column

A	0.26	0.09	0.04	0.00	0.00	0.00	0.00	0.00	0.13	0.04	0.04	0.04	0.00	0.00	0.00	0.09	0.04	0.04	0.13	0.04
C	0.20	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.07	0.13	0.00	0.00	0.00	0.07
D	0.00	0.09	0.00	0.00	0.00	0.27	0.00	0.09	0.00	0.00	0.00	0.09	0.00	0.00	0.09	0.09	0.00	0.18	0.00	0.09
E	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.17	0.33	0.17	0.00	0.00	0.00
F	0.00	0.00	0.00	0.40	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.07	0.07	0.21	0.00	0.07	0.14	0.21	0.07	0.00	0.00	0.07
H	0.00	0.00	0.00	0.20	0.20	0.20	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.00	0.11	0.11	0.00	0.11	0.00	0.22	0.22	0.00	0.00
K	0.00	0.00	0.07	0.00	0.14	0.07	0.07	0.07	0.07	0.00	0.00	0.14	0.07	0.00	0.07	0.07	0.07	0.00	0.00	0.07
L	0.22	0.11	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.00	0.00	0.00
M	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
N	0.04	0.13	0.04	0.00	0.04	0.08	0.00	0.04	0.04	0.08	0.00	0.00	0.04	0.04	0.08	0.04	0.13	0.04	0.04	0.08
P	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17
Q	0.40	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.10	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00
R	0.00	0.23	0.00	0.08	0.00	0.08	0.08	0.00	0.00	0.00	0.00	0.31	0.00	0.08	0.00	0.00	0.08	0.00	0.08	0.00
S	0.08	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.04	0.08	0.00	0.08	0.00	0.04	0.13	0.21	0.13	0.04	0.00	0.04
T	0.12	0.00	0.24	0.00	0.06	0.06	0.00	0.00	0.06	0.00	0.06	0.12	0.06	0.12	0.00	0.06	0.06	0.00	0.00	0.00
V	0.23	0.15	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.15	0.08	0.15	0.00	0.08	0.00	0.00	0.00	0.00
W	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.33	0.17	0.00
Y	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.11	0.11	0.22	0.00	0.11	0.00	0.00

(continued)

Appendix B (*Continued*)

The conditional Probabilities Derived From the 239 Noncleavable Peptides by HIV-1 Protease

4. For the subsite $R_{1'}$: $P^{-}_{1'}(X_{1'}|X_1)$, where X_1, represents the amino acids along the row, and X_1 the amino acid along the column

A	0.26	0.09	0.04	0.00	0.00	0.00	0.00	0.00	0.13	0.04	0.04	0.04	0.00	0.00	0.00	0.09	0.04	0.04	0.13	0.04
C	0.20	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.07	0.13	0.00	0.00	0.00	0.07
D	0.00	0.09	0.00	0.00	0.00	0.27	0.00	0.09	0.00	0.00	0.00	0.09	0.00	0.00	0.09	0.09	0.00	0.18	0.00	0.09
E	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.17	0.33	0.17	0.00	0.00	0.00
F	0.00	0.00	0.20	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
G	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.07	0.07	0.21	0.00	0.07	0.14	0.21	0.07	0.00	0.00	0.07
H	0.00	0.00	0.00	0.20	0.20	0.20	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.00	0.11	0.11	0.00	0.11	0.00	0.22	0.22	0.00	0.00
K	0.00	0.00	0.07	0.00	0.14	0.07	0.07	0.07	0.07	0.00	0.00	0.14	0.07	0.00	0.07	0.07	0.07	0.00	0.00	0.07
L	0.22	0.11	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.00	0.00	0.00
M	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
N	0.04	0.13	0.04	0.00	0.04	0.08	0.00	0.04	0.04	0.08	0.00	0.00	0.04	0.04	0.08	0.04	0.13	0.04	0.04	0.08
P	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17
Q	0.40	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.10	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00
R	0.00	0.21	0.00	0.07	0.00	0.14	0.07	0.00	0.00	0.00	0.00	0.29	0.00	0.07	0.00	0.00	0.07	0.00	0.07	0.00
S	0.08	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.04	0.08	0.00	0.08	0.00	0.04	0.13	0.21	0.13	0.04	0.00	0.04
T	0.06	0.00	0.25	0.00	0.06	0.06	0.00	0.00	0.06	0.00	0.06	0.13	0.06	0.13	0.00	0.06	0.06	0.00	0.00	0.00
V	0.23	0.15	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.15	0.08	0.15	0.00	0.08	0.00	0.00	0.00	0.00
W	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.0	0.00	0.17	0.00	0.00	0.33	0.17	0.00
Y	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.11	0.11	0.22	0.00	0.11	0.00	0.00

5. For the subsite $R_{2'}$: $P^{-}_{2'}(X_{2'}|X_{1'})$, where $X_{2'}$ represents the amino acids along the row, and $X_{1'}$ the amino acid along the column

A	0.23	0.09	0.05	0.00	0.00	0.00	0.00	0.00	0.14	0.05	0.05	0.05	0.00	0.00	0.00	0.09	0.05	0.05	0.14	0.05
C	0.20	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.07	0.13	0.00	0.00	0.00	0.07
D	0.08	0.08	0.00	0.00	0.00	0.25	0.00	0.08	0.00	0.00	0.00	0.08	0.00	0.00	0.08	0.08	0.00	0.17	0.00	0.08
E	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.17	0.33	0.17	0.00	0.00	0.00
F	0.00	0.00	0.20	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
G	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.07	0.07	0.21	0.00	0.07	0.07	0.21	0.07	0.00	0.00	0.07
H	0.00	0.00	0.00	0.20	0.20	0.20	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.00	0.11	0.11	0.00	0.11	0.00	0.22	0.22	0.00	0.00
K	0.00	0.00	0.07	0.00	0.14	0.07	0.07	0.07	0.07	0.00	0.00	0.14	0.07	0.00	0.07	0.07	0.07	0.00	0.00	0.07
L	0.22	0.11	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.00	0.00	0.00
M	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
N	0.04	0.13	0.04	0.00	0.04	0.08	0.00	0.04	0.04	0.08	0.00	0.00	0.04	0.04	0.08	0.04	0.13	0.04	0.04	0.08
P	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17
Q	0.40	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.10	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00
R	0.00	0.21	0.00	0.07	0.00	0.14	0.07	0.00	0.00	0.00	0.00	0.29	0.00	0.07	0.00	0.00	0.07	0.00	0.07	0.00
S	0.08	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.04	0.08	0.00	0.08	0.00	0.04	0.13	0.21	0.13	0.04	0.00	0.04
T	0.06	0.00	0.25	0.00	0.06	0.06	0.00	0.00	0.06	0.00	0.06	0.13	0.06	0.13	0.00	0.06	0.06	0.00	0.00	0.00
V	0.23	0.15	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.15	0.08	0.15	0.00	0.08	0.00	0.00	0.00	0.00
W	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.33	0.17	0.00
Y	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.11	0.11	0.22	0.00	0.11	0.00	0.00

6. For the subsite $R_{3'}$: $P^{-}_{3'}(X_{3'}|X_{2'})$, where $X_{3'}$ represents the amino acids along the row, and X_2, the amino acid along the column

A	0.18	0.09	0.05	0.00	0.00	0.00	0.00	0.00	0.14	0.05	0.05	0.05	0.00	0.00	0.00	0.14	0.05	0.05	0.14	0.05
C	0.19	0.00	0.00	0.13	0.00	0.00	0.00	0.00	0.19	0.00	0.00	0.19	0.00	0.00	0.13	0.13	0.00	0.00	0.00	0.06
D	0.08	0.08	0.00	0.00	0.00	0.25	0.00	0.08	0.00	0.00	0.00	0.08	0.00	0.00	0.08	0.08	0.00	0.17	0.00	0.08
E	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.17	0.33	0.17	0.00	0.00	0.00
F	0.00	0.00	0.20	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
G	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.07	0.07	0.21	0.00	0.07	0.07	0.21	0.07	0.00	0.00	0.07
H	0.00	0.00	0.00	0.20	0.20	0.20	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.00	0.11	0.11	0.00	0.11	0.00	0.22	0.22	0.00	0.00
K	0.00	0.00	0.07	0.00	0.14	0.07	0.07	0.07	0.07	0.00	0.00	0.14	0.07	0.00	0.07	0.07	0.07	0.00	0.00	0.07
L	0.22	0.11	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.00	0.00	0.00
M	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
N	0.04	0.13	0.04	0.00	0.04	0.08	0.00	0.04	0.04	0.08	0.00	0.00	0.04	0.04	0.08	0.04	0.13	0.04	0.04	0.08
P	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17
Q	0.40	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.10	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00
R	0.00	0.15	0.00	0.08	0.00	0.15	0.08	0.00	0.00	0.00	0.00	0.31	0.00	0.08	0.00	0.00	0.08	0.00	0.08	0.00
S	0.08	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.04	0.08	0.00	0.08	0.00	0.04	0.13	0.21	0.13	0.04	0.00	0.04
T	0.06	0.00	0.25	0.00	0.06	0.06	0.00	0.00	0.06	0.00	0.06	0.13	0.06	0.13	0.00	0.06	0.06	0.00	0.00	0.00
V	0.23	0.15	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.15	0.08	0.15	0.00	0.08	0.00	0.00	0.00	0.00
W	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.33	0.17	0.00
Y	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.11	0.11	0.22	0.00	0.11	0.00	0.00

(*continued*)

Appendix B (*Continued*)
The conditional Probabilities Derived From the 239 Noncleavable Peptides by HIV-1 Protease

7. For the subsite $R_{3'}$: $P^{-}_{4'}(X_{4'}|X_{3'})$, where $X_{4'}$ represents the amino acids along the row, and $X_{3'}$ the amino acid along the column

A	0.19	0.10	0.05	0.00	0.00	0.00	0.00	0.00	0.10	0.05	0.05	0.05	0.00	0.00	0.00	0.14	0.05	0.05	0.14	0.05
C	0.20	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.13	0.13	0.00	0.00	0.00	0.07
D	0.08	0.08	0.00	0.00	0.00	0.25	0.00	0.08	0.00	0.00	0.00	0.08	0.00	0.00	0.08	0.08	0.00	0.17	0.00	0.08
E	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.17	0.33	0.17	0.00	0.00	0.00
F	0.00	0.00	0.20	0.40	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.20	0.00	0.00
G	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.07	0.00	0.07	0.07	0.21	0.00	0.07	0.07	0.21	0.07	0.00	0.00	0.07
H	0.00	0.00	0.00	0.20	0.20	0.20	0.00	0.20	0.00	0.00	0.20	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
I	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.11	0.00	0.11	0.00	0.11	0.11	0.00	0.11	0.00	0.22	0.22	0.00	0.00
K	0.00	0.00	0.07	0.00	0.14	0.07	0.07	0.07	0.07	0.00	0.00	0.14	0.07	0.00	0.07	0.07	0.07	0.00	0.00	0.07
L	0.22	0.11	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.11	0.11	0.00	0.00	0.00
M	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00
N	0.04	0.13	0.04	0.00	0.04	0.08	0.00	0.04	0.04	0.08	0.00	0.00	0.04	0.04	0.08	0.04	0.13	0.04	0.04	0.08
P	0.00	0.17	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.33	0.00	0.17
Q	0.40	0.00	0.00	0.00	0.00	0.00	0.10	0.10	0.10	0.00	0.10	0.00	0.00	0.00	0.00	0.10	0.10	0.00	0.00	0.00
R	0.00	0.14	0.00	0.07	0.00	0.14	0.07	0.00	0.00	0.07	0.00	0.29	0.00	0.07	0.00	0.00	0.07	0.00	0.07	0.00
S	0.08	0.00	0.08	0.00	0.00	0.00	0.00	0.04	0.04	0.08	0.00	0.08	0.00	0.04	0.12	0.20	0.12	0.08	0.00	0.04
T	0.06	0.00	0.25	0.00	0.06	0.06	0.00	0.00	0.06	0.00	0.06	0.13	0.06	0.13	0.00	0.06	0.06	0.00	0.00	0.00
V	0.23	0.15	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.15	0.08	0.15	0.00	0.08	0.00	0.00	0.00	0.00
W	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.00	0.00	0.00	0.00	0.17	0.00	0.00	0.33	0.17	0.00
Y	0.00	0.11	0.00	0.00	0.00	0.11	0.00	0.00	0.11	0.00	0.00	0.00	0.11	0.11	0.11	0.22	0.00	0.11	0.00	0.00