Journal of Protein Chemistry, Vol. 12, No. 3, 1993

Predicting Cleavability of Peptide Sequences by HIV Protease via Correlation-Angle Approach

James J. Chou[1,2]

Received December 17, 1992

In designing HIV protease inhibitors as potential drugs for AIDS therapy, knowledge about what peptide sequences in polyproteins are cleavable by HIV proteases is very useful. In this article, based on the formulation that any octapeptide can be uniquely expressed as a 160-dimensional vector and the principle that the similarity of any two such vectors is associated with their correlation angle, a new method is proposed to predict the cleavability of a peptide sequence by HIV-1 and HIV-2 proteases. The average predicted accuracy the new method for the 105 peptide sequences whose cleavability by HIV-1 protease is known is 96/105 = 9.14%, which is about 8% higher than that by the existing method for the same set of data. A considerably high rate of correct prediction was also obtained when the new method was used to predict the HIV-2 protease-cleaved sites in some proteins.

KEY WORDS: HIV-1 protease; HIV-2 protease; specificity; 160-D space; normal distribution.

1. INTRODUCTION

Since the discovery of the disease a decade ago, acquired immunodeficiency syndrome (AIDS) has become a synonym of terror. The magnitude of the mounting AIDS problem is sobering. Since the initial clinical reports in 1981, over 210,000 cases of AIDS have been diagnosed in the United States, and nearly half of these patients have died as a result of the disease. Globally, the World Health Organization estimates that 6–8 million people, or perhaps as many as 1 in every 400 adults, are currently infected with HIV. A new report on AIDS by the World Health Organization predicts a far grimmer future than previously forecast, warning that more than 100 million people worldwide could be infected by the end of the decade. As claimed by some experts, "AIDS is a global epidemic that is heading out of control." AIDS has galvanized the concern and efforts of physicians, scientists, and even the lay public alike. Actually, it has presented to scientists of all areas a significant challenge (i.e., how to provide any useful knowledge and technology at all that will lead to finding effective drugs against AIDS).

It has been clearly identified that human immunodeficiency virus (HIV) is the primary cause of AIDS (Barré-Sinoussi *et al.*, 1983; Gallo *et al.*, 1984). Therefore, a key step against AIDS is how to suppress HIV. It has been known that the replication of HIV is accompanied by the process in which some high molecular weight polyproteins are cleaved by a specific enzyme called HIV protease. This processing is indispensable to the viral reproduction (Kohl *et al.*, 1988; Hellen *et al.*, 1989; Wlodawer *et al.*, 1989). Therefore, one of the effective avenues in suppressing the growth of HIV is to inhibit the HIV protease. Many efforts have been made in order to find specific inhibitors to inactivate HIV-protease (Putney, 1992). In this regard, information about the HIV protease cleavage sites in polyproteins is very useful in refining our understanding of the specificity. Moreover, the knowledge thus acquired can play a guiding role for designing HIV protease inhibitors as potential drugs

[1]Department of Physics, University of Michigan, Ann Arbor, MI 48104, USA.
[2]Present address: 4416 Woodhaven Drive, Kalamazoo, MI 49008, USA.

0277-8033/93/0600-0291$07.00/0

for AIDS therapy (Hellen *et al.*, 1989; Henderson *et al.*, 1988). Consequently, it is very useful to develop a method to predict the cleavability of a peptide sequence by HIV protease.

Recently, based on a series of sequences surrounding HIV protease cleavage sites in proteins, a cumulative specificity model was proposed (Poorman *et al.*, 1991) to characterize the substrate specificity of HIV protease. According to their model, the moiety of susceptible sites in polyproteins is usually an octapeptide, although it may occasionally be a heptapeptide. Furthermore, if the positions of the eight amino acids of an octapeptide are subsequently expressed as P_4, P_3, P_2, P_1, $P_{1'}$, $P_{2'}$, $P_{3'}$, $P_{4'}$, then the bond to be cleaved by the enzyme, the so-called scissile bond, is the one between P_1 and $P_{1'}$. According to their model, an octapeptide is characterized by an h function, which is actually a multiplication of some parameters derived from a set of peptides known cleavable by HIV protease. The prediction of cleavability for a given peptide is based on the following criterion: if its h value is greater than h^*, the so-called "cutoff" value or critical value, the peptide is assumed to be cleavable by HIV protease; otherwise, it is not. Accordingly, their method can also be termed as h function method. For the case of HIV-1 protease, according to their report, the rate of correct prediction for 74 Therefore, one of the effective avenues in suppressing the growth of HIV is to inhibit the HIV protease. Many efforts have been made in order to find specific inhibitors to inactivate HIV-protease (Putney, 1992). In this regard, information about the HIV protease cleavage sites in polyproteins is very useful in refining our understanding of the specificity. Moreover, the knowledge thus acquired can play a guiding role for designing HIV protease inhibitors as potential drugs for AIDS therapy (Henderson *et al.*, 1988; Hellen *et al.*, 1989). Consequently, it is very useful to develop a method to predict the cleavability of a peptide sequence by HIV protease.

2. METHOD

Any octapeptide x can be expressed as

$$x \Leftrightarrow X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'} \tag{1}$$

where X_i ($i = 4, 3, 2, 1, 1', 2', 3', 4'$) represent the amino acid occupying subsite P_i. Since each of its eight subsites can be occupied by any of 20 amino acids, the octapeptide x can be uniquely defined as a $20 \times 8 = 160$-D (-dimensional) vector, as formulated as follows:

$$\begin{aligned}\mathbf{\Phi}(x) = \{&\ldots, \phi_4^i(x), \ldots, \phi_3^i(x), \ldots, \phi_2^i(x),\\ &\ldots, \phi_1^i(x), \ldots, \phi_{1'}^i(x), \ldots, \phi_{2'}^i(x),\\ &\ldots, \phi_{3'}^i(x), \ldots, \phi_{4'}^i(x), \ldots\}\\ &(i = 1, 2, \ldots, 20)\end{aligned} \tag{2}$$

where $\phi_4^i(x) = 1$ if position P_4 is occupied by the ith amino acid, otherwise $\phi_4^i(x) = 0$, and so forth. Here, the amino acids are numbered according to the alphabetic order of their one-letter code—that is, $i = 1, 2, \ldots, 20$ for A (alanine), C (cysteine), ..., Y (tyrosine), respectively. For example, if an octapeptide x = ACACYYYY, then its characteristic vector in the 160-D space is

$$\mathbf{\Phi}(\text{ACACYYYY}) = \left\{\begin{array}{l}1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,\\ 0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,\\ 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,\\ 0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,\\ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,\\ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,\\ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,\\ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1\end{array}\right\} \tag{3}$$

Occasionally, one has to deal with heptapeptides in which the position P_4 or $P_{4'}$ is not occupied by any amino acid (Poorman *et al.*, 1991). For cases like that, just put $\phi_4^i(x) = 0$ or $\phi_{4'}^i(x) = 0$ for all of $i = 1, 2, \ldots, 20$, and all the formulation described here is valid for hepaptides as well. On the other hand, the norm of the cleavable peptide sequences by HIV-1 protease can also be defined in the same 160-D space, as given by

$$\begin{aligned}\mathbf{\Phi}(\text{H-1}) = \{&\ldots, \phi_4^i(\text{H-1}), \ldots, \phi_3^i(\text{H-1}), \ldots, \phi_2^i(\text{H-1}),\\ &\ldots, \phi_1^i(\text{H-1}), \ldots, \phi_{1'}^i(\text{H-1}), \ldots, \phi_{2'}^i(\text{H-1}),\\ &\ldots, \phi_{3'}^i(\text{H-1}), \ldots, \phi_{4'}^i(\text{H-1}), \ldots\}\end{aligned} \tag{4}$$

where H-1 represents the norm of octapeptides cleavable by HIV-1 protease and its components in the 160-D space are given by

$$\begin{aligned}&\phi_j^i(\text{H-1}) = p_j^i(\text{H-1}) - \mu^i\\ &(i = 1, 2, \ldots, 20; j = 4, 3, \ldots, 3', 4')\end{aligned} \tag{5}$$

where p_j^i(H-1) is the frequency of amino acid i occurring in the position P_j as will be derived later from a training set consisting of cleavable oligopeptides by HIV-1 protease, and μ^i the abundance of amino acid i in globular proteins (Nakashima *et al.*, 1986).

The components thus defined for the norm $\mathbf{\Phi}$(H-1) are justifed by the fact that if amino acid i in position P_j is critical to the cleavage specificity, its frequency p_j^i(H-1) in $\mathbf{\Phi}$(H-1) should be significantly greater than its abundance, μ^i, in globular proteins, and hence the corresponding component ϕ_j^i(H-1) is positive. If amino acid i in position P_j plays an unfavorable or immaterial role to the specificity, then the corresponding component ϕ_j^i(H-1) would become negative or zero, respectively. Therefore, unlike Eq. (3) which represents a clear-cut octapeptide in the 160-D space as reflected by either 0 or 1 in its components, the HIV-1 norm as defined by Eqs. (4) and (5) reflects the composition probabilities of various cleavable peptides.

Similarly, the standard vector representing the norm of peptide sequences cleavable by HIV-2 protease can be expressed by

$$\begin{aligned}\mathbf{\Phi}(\text{H-2}) = \{&\ldots, \phi_4^i(\text{H-2}), \ldots, \phi_3^i(\text{H-2}), \ldots, \phi_2^i(\text{H-2}),\\ &\ldots, \phi_1^i(\text{H-2}), \ldots, \phi_{1'}^i(\text{H-2}), \ldots, \phi_{2'}^i(\text{H-2}),\\ &\ldots, \phi_{3'}^i(\text{H-2}), \ldots, \phi_{4'}^i(\text{H-2}), \ldots\} \end{aligned} \tag{6}$$

where H-2 represents the norm of octapeptides cleavable by HIV-2 protease and its components in the 160-D space are given by

$$\begin{gathered}\phi_j^i(\text{H-2}) = p_j^i(\text{H-2}) - \mu^i \\ (i = 1, 2, \ldots, 20; j = 4, 3, \ldots, 3', 4')\end{gathered} \tag{7}$$

where p_j^i(H-2) is the frequency of amino acid i occurring in the position P_j for the norm of peptides cleavable by HIV-2 protease.

According to the Cauchy–Schwartz–Buniakowsky inequality, for any two arbitrary sets of numbers $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$, we have

$$\left(\sum_{k=1}^{n} a_k b_k\right)^2 \leq \left(\sum_{k=1}^{n} a_k^2\right)\left(\sum_{k=1}^{n} b_k^2\right) \tag{8}$$

The equality holds if, and only if, the sequences $a_1, a_2, \ldots, a_n$ and $b_1, b_2, \ldots, b_n$ are proportional. Thus, the correlation angles of the vector $\mathbf{\Phi}(x)$ with $\mathbf{\Phi}$(H-1) and $\mathbf{\Phi}$(H-2) can be defined as follows:

$$\begin{aligned}\Theta_{\text{H-1}}(x) &= \cos^{-1}\left\{\frac{\mathbf{\Phi}(x)\cdot\mathbf{\Phi}(\text{H-1})}{|\mathbf{\Phi}(x)|\,|\mathbf{\Phi}(\text{H-1})|}\right\}\\ \Theta_{\text{H-2}}(x) &= \cos^{-1}\left\{\frac{\mathbf{\Phi}(x)\cdot\mathbf{\Phi}(\text{H-2})}{|\mathbf{\Phi}(x)|\,|\mathbf{\Phi}(\text{H-2})|}\right\}\end{aligned} \tag{9}$$

or

$$\begin{aligned}&\Theta_{\text{H-1}}(x)\\ &= \cos^{-1}\left\{\frac{\sum_{j=1}^{4}\sum_{i=1}^{20}\phi_j^i(x)\phi_j^i(\text{H-1})}{\{[\sum_{j=1}^{4}\sum_{i=1}^{20}\phi_j^i(x)^2][\sum_{j=1}^{4}\sum_{i=1}^{20}\phi_j^i(\text{H-1})^2]\}^{1/2}}\right\}\\ &\Theta_{\text{H-2}}(x)\\ &= \cos^{-1}\left\{\frac{\sum_{j=1}^{4}\sum_{i=1}^{20}\phi_j^i(x)\phi_j^i(\text{H-2})}{\{[\sum_{j=1}^{4}\sum_{i=1}^{20}\phi_j^i(x)^2][\sum_{j=1}^{4}\sum_{i=1}^{20}\phi_j^i(\text{H-2})^2]\}^{1/2}}\right\}\end{aligned} \tag{10}$$

where $\Theta_{\text{H-1}}(x)$ is the correlation angle of the vector $\mathbf{\Phi}(x)$ for the octapeptide x with the standard vector $\mathbf{\Phi}$(H-1) representing the norm of peptide sequences cleavable by HIV-1 protease, and $\Theta_{\text{H-2}}(x)$ is the correlation angle of the vector $\mathbf{\Phi}(x)$ with the vector $\mathbf{\Phi}$(H-2) representing the norm of peptide sequences cleavable by HIV-2 protease. Define

$$\begin{aligned}\Delta\Theta_{\text{H-1}}(x) &= \Theta_{\text{H-1}}^* - \Theta_{\text{H-1}}(x)\\ \Delta\Theta_{\text{H-2}}(x) &= \Theta_{\text{H-2}}^* - \Theta_{\text{H-2}}(x)\end{aligned} \tag{11}$$

where the parameters $\Theta_{\text{H-1}}^*$ and $\Theta_{\text{H-2}}^*$ are the upper limits of correlation angle for the peptide sequences cleavable by HIV-1 and HIV-2 proteases, respectively, and they can be determined through an optimization procedure as will be illustrated later.

Thus, whether an octapeptide x can be cleaved by HIV-1 or HIV-2 protease will depend on the value of $\Theta_{\text{H-1}}(x)$ or $\Theta_{\text{H-2}}(x)$, as can be formulated by the following eqs. (12) and (13):

$$\begin{cases}\text{An octapeptide } x \text{ can be cleaved by HIV-1 protease,} & \text{if } \Delta\Theta_{\text{H-1}}(x) \geq 0\\ \text{An octapeptide } x \text{ cannot be cleaved by HIV-1 protease,} & \text{if } \Delta\Theta_{\text{H-1}}(x) < 0\end{cases} \tag{12}$$

$$\begin{cases}\text{An octapeptide } x \text{ can be cleaved by HIV-2 protease,} & \text{if } \Delta\Theta_{\text{H-2}}(x) \geq 0\\ \text{An octapeptide } x \text{ cannot be cleaved by HIV-2 protease,} & \text{if } \Delta\Theta_{\text{H-2}}(x) < 0\end{cases} \tag{13}$$

The physical implication of Eqs. (12) and (13) can be further illustrated as follows. As is well known, the smaller the projection angle between two vectors, the larger their mutual projection, and hence the stronger their similarity. Thus, the vector $\mathbf{\Phi}(x)$ with $\Delta\Theta_{\text{H-1}}(x) \geq 0$ (or $\Delta\Theta_{\text{H-2}}(x) \geq 0$) is closer to the norm of the cleavable peptides than the vector $\mathbf{\Phi}(x)$ with $\Delta\Theta_{\text{H-1}}(x) < 0$ (or $\Delta\Theta_{\text{H-2}}(x) < 0$), and hence is more likely to be cleavable by HIV-1 (or HIV-2) protease.

In other words, the current approach provides a quantitative description for the similarity between any given peptide sequence and the standard cleavable peptide in terms of the corresponding projection angle as defined in Eq. (10).

It should be pointed out that this method, like the h function method, depends on the independent-subsite specificity assumption (i.e., the "best" amino acid at position P_1 is completely independent of amino acid present at P_2 and $P_{1'}$). In many cases this is a valid first approximation according to Schechter and Berger (1967).

3. RESULTS AND DISCUSSION

According to Eqs. (12) and (13), to judge whether a peptide x can be cleaved by HIV-1 or HIV-2 protease, we have to first calculate $\Theta_{\text{H-1}}(x)$ or $\Theta_{\text{H-2}}(x)$, the correlation angle of $\mathbf{\Phi}(x)$ with $\mathbf{\Phi}$(H-1) or $\mathbf{\Phi}$(H-2), respectively. In order to realize that, we have to find ψ_j^i(H-1) ($i = 1, 2, \ldots, 20; j = 4, 3, \ldots, 3', 4'$) or ψ_j^i(H-2) ($i = 1, 2, \ldots, 20; j = 4, 3, \ldots, 3', 4'$), the 160 components for the standard vector $\mathbf{\Phi}$(H-1) or $\mathbf{\Phi}$(H-2), respectively [see Eqs. (4) and (6)]. Because the abundance for each of the 20 amino acids in globular proteins is known (Nakashima *et al.*, 1986), if we can find p_j^i(H-1) or p_j^i(H-2), we can immediately obtain ψ_j^i(H-1) or ψ_j^i(H-2) by means of Eqs. (5) or (7), respectively. Actually, p_j^i(H-1) or p_j^i(H-2) can be derived from a set of octapeptides known to be cleavable by HIV-1 or HIV-2 protease. Such a set of data is usually termed as "training set" or "development set." Below, let us first consider the case associated with HIV-1 protease.

3.1. HIV-1 Protease

In order to compare the predicted results by means of the current correlation angle method with those by the h probability method, we should use the same set of training data. The following 40 peptide sequences, of which 38 are octapeptides and 2 heptapeptides, have been found cleavable by HIV-1 protease. In Eq. (14) the arrow indicates the scissile bond, which is between the positions P_1 and $P_{1'}$. The above 40 peptide sequences were used by the h probability method (Poorman *et al.*, 1991) as a training set for HIV-1 protease. Using the same training set [i.e., Eq. (14)], we can derive p_j^i(H-1) ($i = 1, 2, \ldots, 20; j = 4, 3, \ldots, 3', 4'$), which, together with the abundance of amino acids, are listed in Table I.

P_4	P_3	P_2	P_1	—	$P_{1'}$	$P_{2'}$	$P_{3'}$	$P_{4'}$
				⇓				
T	Q	I	M	—	F	E	T	F
G	Q	V	N	—	Y	E	E	F
P	F	I	F	—	E	E	E	P
S	F	N	F	—	P	Q	I	T
D	T	V	L	—	E	E	M	S
A	R	V	L	—	A	E	A	M
A	E	E	L	—	A	E	I	F
S	L	N	L	—	R	E	T	N
A	T	I	M	—	M	Q	R	G
A	E	C	F	—	R	I	F	D
D	Q	I	L	—	I	E	I	C
D	D	L	F	—	F	E	A	D
Y	E	E	F	—	V	Q	M	M
P	I	V	G	—	A	E	T	F
T	L	N	F	—	P	I	S	P
R	E	A	F	—	R	V	F	D
A	E	T	F	—	Y	V	D	K
A	Q	T	F	—	Y	V	N	L
P	T	L	L	—	T	E	A	P
S	F	I	G	—	M	E	S	A
D	A	I	N	—	T	E	F	K
Q	I	T	L	—	W	Q	R	P
E	L	E	F	—	P	E	G	G
	A	N	L	—	A	E	E	A
S	Q	N	Y	—	P	I	V	Q
P	G	N	F	—	L	Q	S	R
K	L	V	F	—	F	A	E	
G	D	A	L	—	L	E	R	N
K	E	L	Y	—	P	L	T	S
R	Q	A	N	—	F	L	G	K
S	R	S	L	—	Y	A	S	S
A	E	A	M	—	S	Q	V	T
R	K	I	L	—	F	L	D	G
G	S	H	L	—	V	E	A	L
G	G	V	Y	—	A	T	R	S
F	R	S	G	—	V	E	T	T
V	E	V	A	—	E	E	E	E
L	P	V	N	—	G	E	F	S
E	T	T	A	—	L	V	C	D
H	L	V	E	—	A	L	Y	L

(14)

Based on Table I, for any given octapeptide we can calculate its projection angle $\Theta_{\text{H-1}}(x)$ with the standard cleavable vector $\mathbf{\Phi}$(H-1) for HIV-1 protease according to Eqs. (5) and (10). Thus, according to Eq. (11), once the value of $\Theta^*_{\text{H-1}}$ is determined, $\Delta\Theta_{\text{H-1}}(x)$ is uniquely defined. To realize this, let us apply an optimization procedure, which is actually a

Table I. Values of p_j^i Derived from Eq. (14) for Standard Vector Ψ(H-1)

Amino acid		Abundance[a]	Probability of amino acid i at each of eight positions							
Index	Code	μ^i	p_4^i	p_3^i	p_2^i	p_1^i	$p_{1'}^i$	$p_{2'}^i$	$p_{3'}^i$	$p_{4'}^i$
1	A	0.087	7/39	2/40	4/40	2/40	6/40	2/40	4/40	2/39
2	C	0.016	0/39	0/40	1/40	0/40	0/40	0/40	1/40	1/39
3	D	0.057	4/39	2/40	0/40	0/40	0/40	0/40	2/40	4/39
4	E	0.064	2/39	8/40	3/40	1/40	3/40	20/40	5/40	1/39
5	F	0.039	1/39	3/40	0/40	12/40	5/40	0/40	4/40	4/39
6	G	0.078	4/39	2/40	0/40	3/40	1/40	0/40	2/40	3/39
7	H	0.022	1/39	0/40	1/40	0/40	0/40	0/40	0/40	0/39
8	I	0.052	0/39	2/40	7/40	0/40	1/40	3/40	3/40	0/39
9	K	0.068	2/39	1/40	0/40	0/40	0/40	0/40	0/40	3/39
10	L	0.082	1/39	5/40	3/40	12/40	3/40	4/40	0/40	3/39
11	M	0.021	0/39	0/40	0/40	3/40	2/40	0/40	2/40	2/39
12	N	0.044	0/39	0/40	6/40	4/40	0/40	0/40	1/40	2/39
13	P	0.045	4/39	1/40	0/40	0/40	5/40	0/40	0/40	4/39
14	Q	0.039	1/39	6/40	0/40	0/40	0/40	6/40	0/40	1/39
15	R	0.048	3/39	3/40	0/40	0/40	3/40	0/40	4/40	1/39
16	S	0.066	5/39	1/40	2/40	0/40	1/40	0/40	4/40	5/39
17	T	0.058	2/39	4/40	4/40	0/40	2/40	1/40	5/40	3/39
18	V	0.070	1/39	0/40	9/40	0/40	3/40	4/40	2/40	0/39
19	W	0.012	0/39	0/40	0/40	0/40	1/40	0/40	0/40	0/39
20	Y	0.033	1/39	0/40	0/40	3/40	4/40	0/40	1/40	0/39

[a] The values of amino acid abundance in globular proteins are taken from Nakashima *et al.* (1986).

compromise between overpredicting for a set of noncleavable peptides and underpredicting for a set of cleavable peptides. It is obvious from Eqs. (11) and (12) that if $\Theta_{H\text{-}1}$ is too large, then some noncleavable oligopeptides by the enzyme will be overpredicted as cleavable. On the other hand, if $\Theta^*_{H\text{-}1}$ is too small, some cleavable oligopeptides will be underpredicted as noncleavable. Therefore, to find the optimal value for $\Theta^*_{H\text{-}1}$, two types of training data are needed: one is of cleavable peptide, and the other is of noncleavable peptide. We already have the training data for the former (i.e. the 40 oligopeptides listed in Table II). The training data for the latter can be obtained as follows. It has been found that only three octapeptides in calmodulin, i.e. GQVN-YEEF, YEEF-VQMM, and REAF-RVFD, are cleavable by HIV-1 protease. No any other HIV-1 protease cleavage sites whatsoever were detected for calmodulin even after it was completely denatured (Poorman *et al.*, 1991). This protein contains 148 residues, and hence can provide $148 - 7 = 141$ cases for oligopeptides. Excluding the above three cleavable peptides, we still have $141 - 3 = 138$ octapeptides, which constitute a set of non-cleavable peptides and can be used as a benchmark for checking overpredicted results. The process for finding the optimal $\Theta^*_{H\text{-}1}$ is illustrated in Fig. 1, where the number of correct prediction vs. ϕ_1 for the 40 cleavable oligopeptides is plotted by the dotted line, and that for the 138 noncleavable peptides is plotted by the dashed line. The solid line indicates the average of the two rates. As we can see from Fig. 1, for the 40 cleavable oligopeptides, the greater the $\Theta^*_{H\text{-}1}$, the higher the rate of correct predictions; while for the 138 noncleavable octapeptides, the situation is just the opposite. As a combined result of these two opposite changes with $\Theta_{H\text{-}1}$, there is a peak at $\Theta^*_{H\text{-}1} = 82.4°$ for the average of these two rates. Such a value is taken as the optimal value for $\Theta^*_{H\text{-}1}$ because it leads to the highest average rate of correct predictions.

The predicted results for the 40 oligopeptides in the training set for HIV-1 protease are given in Table II. It is shown in that table that there are only three octapeptides (i.e., RQAN-FLGK, ETTA-LVCD, and HLVE-ALYL), whose $\Delta\Theta_{H\text{-}1}(x) < 0$. This means that, except these three, all the other oligopeptides are correctly predicted as cleavable because the deviation of their characteristic vectors $\Phi(x)$ from the standard cleavable vector for HIV-1 protease Φ(H-1) is within the upper limit $\Theta^*_{H\text{-}1}$ However, according to the h function method (Poorman *et al.*, 1991) in which the criterion for a cleavable peptide sequence is that its h value must $\geq h^*0.13$, we find that there are eight incorrect prediction results. Consequently, the rate of correction prediction is $37/40 = 92.5\%$ by using the

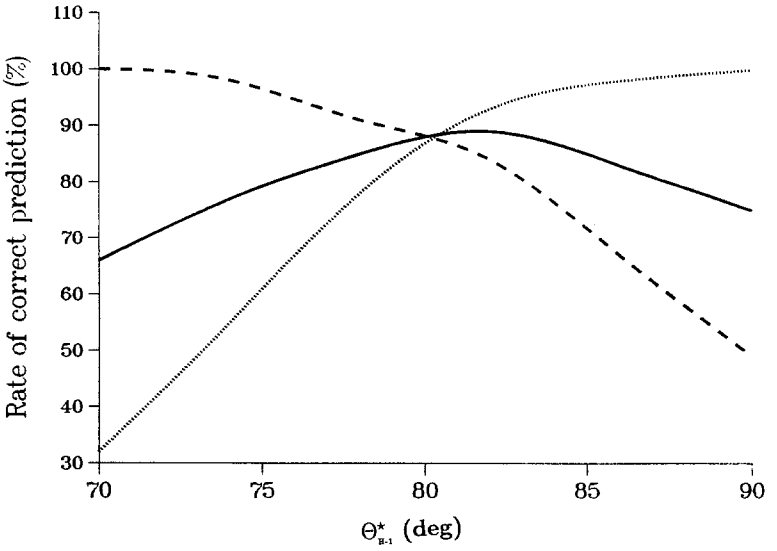

Fig. 1. Plot to show the dependence of the rate of corrected prediction on the critical angle $\Theta^*_{H\text{-}I}$: the rate of corrected prediction vs. $\Theta^*_{H\text{-}I}$ for the cleavable training set (i.e., the 40 oligopeptides listed in Table II) is depicted by the dotted line; the rate of corrected prediction vs. $\Theta^*_{H\text{-}I}$ for the noncleavable training set (i.e., the 138 noncleavable octapeptides in calmodulin) is depicted by the dashed line; and an average of these two rates is depicted by the solid line. As shown by the plot, the larger the $\Theta^*_{H\text{-}I}$, the higher the rate of corrected prediction for the cleavable peptide set (see dotted line), but the lower the rate of corrected prediction for the noncleavable peptide set (see dashed line). As a consequence, their average reaches a maximum (see solid line) at $\Theta^*_{H\text{-}I} = 82.4°$, which is taken as the optimal parameter to predict the cleavability by HIV-1 protease according to Eq. (11).

current projection angle method, while only 32/40 = 80.0% by the *h* function method. This means that the new method has improved the self-consistency in dealing with the training data set.

The prediction results for the 34 octapeptides in a series of wild-type and mutant proteins (Partin *et al.*, 1990) are listed in Table III. Note that although the octapeptides listed here are taken from Poorman *et al.* (1991), any duplicates in either themselves or to the octapeptides in the training set of Table II should be excluded. This is because in either the *h* function method or the projection angle method, the sequence of an octapeptide is the sole input in predicting its cleavability by HIV-protease. Therefore, the total countable testing octapeptides in Table III should be 34 rather than 42. These 34 oligopeptides are outside the training set and hence then can be regarded as an independent testing set. It is shown in Table III that, for these 34 peptides, both methods have 30 correct predicted results (i.e., for the testing set selected by the authors of the *h* function method the rate of correct prediction for both methods is the same, equal to 30/34 = 88.2%.

Bláha *et al.* (1991) have synthesized some analogs of an HIV-1 protease substrate and observed their cleavability. A prediction for these peptide sequences by the current method and that by the *h* function method are listed in Table IV. As shown from the table, for three of the 11 oligopeptides, the predicted results are incorrect if the *h* function is used. But if using the current projection method, only two results are incorrectly predicted (i.e., a slightly better result is obtained).

However, if the comparison for independent testing data is extended to cover more oligopeptides, a result in favor of the projection angle method would become apparent. According to the recent report by Griffiths *et al.* (1992), the 20 oligopeptides listed in Table V are cleavable by HIV-1 protease. The predicted results for these oligopeptides by the current method and that by the *h* function method are also given in Table V. As shown from the table, in two of 20 events, the results were incorrectly predicted by the *h* function method, meaning the rate of correct prediction was 18/20 = 90.0%. However, if using the current sequence-coupled method, none of them was incorrectly predicted (i.e., a rate of correct prediction of 20/20 = 100%!).

The average accuracy can be obtained by combining the data in Tables II–V, which turns out to be 96/105 = 91.4% for the current method but only 88/105 = 83.6% for the *h* function method. This indicates that for the same set of data the average inaccuracy of the projection angle method is about 8% higher than that of the *h* function method.

According to statistical mathematics, a normal distribution should approximately satisfy the following empirical rule (Mendenhall *et al.*, 1986):

$$\begin{cases} M \pm S & \text{contains approximately 66\% of the predictions} \\ M \pm 2S & \text{contains approximately 95\% of the predictions} \\ M \pm 3S & \text{contains almost all of the predictions} \end{cases} \quad (15)$$

Table II. Predicted Results of 40 Peptide Sequences in Training Set for HIV-1 Protease

P_4	P_3	P_2	P_1	—	$P_{1'}$	$P_{2'}$	$P_{3'}$	$P_{4'}$	$\Delta\Theta_{H\text{-}2}(x)$ (deg)[a]	h[b]	Protein
	T	Q	I	M	—	F	E	T	F	16.5	0.97	Actin
	G	Q	V	N	—	Y	E	E	F	17.6	0.96	Calmodulin
	P	F	I	F	—	E	E	E	P	19.6	0.96	pro-IL1-β
	S	F	N	F	—	P	Q	I	T	10.2	0.92	*pol*
	D	T	V	L	—	E	E	M	S	18.3	0.90	Autolysis
	A	R	V	L	—	A	E	A	M	19.3	0.89	*gag*
	A	E	E	L	—	A	E	I	F	19.6	0.89	Troponin C
	S	L	N	L	—	R	E	T	N	17.4	0.87	Vimentin
	A	T	I	M	—	M	Q	R	G	5.1	0.82	*gag*
	A	E	C	F	—	R	I	F	D	9.1	0.82	Troponin C
	D	Q	I	L	—	I	E	I	C	16.7	0.81	Autolysis
	D	D	L	F	—	F	E	A	D	14.9	0.77	pro-IL1-β
	Y	E	E	F	—	V	Q	M	M	7.0	0.75	Calmodulin
	P	I	V	G	—	A	E	T	F	13.9	0.75	*pol*
	T	L	N	F	—	P	I	S	P	7.6	0.74	*pol*
	R	E	A	F	—	R	V	F	D	7.7	0.72	Calmodulin
	A	E	T	F	—	Y	V	D	K	8.4	0.68	*plo*
	A	Q	T	F	—	Y	V	N	L	7.1	0.58	*pol*
	P	T	L	L	—	T	E	A	P	13.2	0.57	Actin
	S	F	I	G	—	M	E	S	A	9.8	0.53	Actin
	D	A	I	N	—	T	E	F	K	9.9	0.47	Vimentin
	Q	I	T	L	—	W	Q	R	P	4.5	0.46	Autolysis
	E	L	E	F	—	P	E	G	G	12.6	0.46	PE664E
		A	N	L	—	A	E	E	A	13.2	0.39	PE40
	S	Q	N	Y	—	P	I	V	Q	2.3	0.38	*gag*
	P	G	N	F	—	L	Q	S	R	5.4	0.38	*gag*
	K	L	V	F	—	F	A	E		6.4	0.38	AAP
	G	D	A	L	—	L	E	R	N	11.2	0.33	PE40
	K	E	L	Y	—	P	L	T	S	2.0	0.28	*gag*
	R	Q	A	N	—	F	L	G	K	−0.2[c]	0.21	*gag*
	S	R	S	L	—	Y	A	S	S	2.9	0.20	Vimentin
	A	E	A	M	—	S	Q	V	T	1.6	0.17	*gag*
	R	K	I	L	—	F	L	D	G	3.1	0.12[d]	*pol*
	G	S	H	L	—	V	E	A	L	9.0	0.10[d]	Insulin
	G	G	V	Y	—	A	T	R	S	0.9	0.10[d]	Vimentin
	F	R	S	G	—	V	E	T	T	5.6	0.09[d]	*gag*
	V	E	V	A	—	E	E	E	E	9.7	0.08[d]	AAP
	L	P	V	N	—	G	E	F	S	8.7	0.08[d]	AAP
	E	T	T	A	—	L	V	C	D	−4.8[c]	0.03[d]	Actin
	H	L	V	E	—	A	L	Y	L	−1.8[c]	0.02[d]	Insulin

[a] See Eq. (11), where $\Theta^*_{H\text{-}1} = 82.4°$ is derived through the optimization procedure as described in the text.
[b] According to the h function method (Poorman *et al.*, 1991), an octapeptide can be cleaved by HIV-1 protease when its $h \geqslant h^* = 0.13$. Otherwise, it cannot be cleaved by the enzyme.
[c] Incorrect prediction by the projection angle method.
[d] Incorrect prediction by the h function method.

where M and S represent mean and standard deviation of the predicted quantity, respectively. Now let us see what distribution we have for the 105 predicted results. Based on the data listed in Tables II–V, it is found that for $\Delta\Theta_{H\text{-}1}(x)$ we have:

$$\begin{cases} M \pm S & \text{contains } 70/105 \approx 66\% \text{ of the predictions} \\ M \pm 2S & \text{contains } 99/105 \approx 95\% \text{ of the predictions} \\ M \pm 3S & \text{contains } 105/101 = 100\% \text{ (i.e., all of the predictions} \end{cases} \tag{16}$$

Table III. Predicted Results for 34 Octapeptides[a] in Series of Wild-Type and Mutant Proteins (Partin *et al.*, 1990)

P_4	P_3	P_2	P_1	—	$P_{1'}$	$P_{2'}$	$P_{3'}$	$P_{4'}$	$\Delta\Theta_{H\text{-}I}(x)$(deg)[b]	h[c]	Experimental[d]
R	Q	N	Y	—	P	I	V	Q	1.4[e]	0.34[f]	−
S	Q	K	Y	—	P	I	V	Q	−2.2	0.03	−
S	Q	Q	Y	—	P	I	V	Q	−1.4	0.02	−
S	Q	N	S	—	P	I	V	Q	−0.5	0.03	−
S	Q	N	P	—	P	I	V	Q	0.1[e]	0.04	−
S	Q	N	Y	—	P	K	V	Q	−0.1	0.04	−
T	Q	N	Y	—	P	I	V	Q	0.5	0.22	+
S	N	N	Y	—	P	I	V	Q	−1.7[e]	0.02[f]	+
S	K	N	Y	—	P	I	V	Q	−1.6[e]	0.06[f]	+
S	Q	N	F	—	P	I	V	Q	7.9	0.68	+
S	Q	N	Y	—	A	I	V	Q	1.8	0.28	+
S	Q	N	Y	—	L	I	V	Q	0.1	0.22	+
S	Q	N	Y	—	T	I	V	Q	0.1	0.16	+
S	Q	N	Y	—	P	V	V	Q	2.4	0.38	+
S	Q	N	Y	—	P	I	I	Q	3.4	0.56	+
S	Q	N	Y	—	P	I	E	Q	4.3	0.63	+
S	Q	N	Y	—	P	I	V	P	4.1	0.68	+
S	Q	N	Y	—	P	I	V	E	1.6	0.28	+
S	F	N	F	—	P	Q	I	T	10.2	0.92	+
T	F	N	F	—	P	Q	I	T	8.4	0.84	+
Y	F	N	F	—	P	Q	I	T	8.4	0.82	+
S	C	N	F	—	P	Q	I	T	8.8	0.75	+
S	Y	N	F	—	P	Q	I	T	8.4	0.53	+
S	F	T	F	—	P	Q	I	T	8.5	0.85	+
S	F	Y	F	—	P	Q	I	T	6.6	0.39	+
S	F	N	S	—	P	Q	I	T	1.8	0.12[f]	+
S	F	N	Y	—	P	Q	I	T	4.5	0.77	+
S	F	N	Y	—	G	Q	I	T	6.7	0.57	+
S	F	N	Y	—	L	Q	I	T	7.9	0.79	+
S	F	N	Y	—	P	P	I	T	6.1	0.29	+
S	F	N	Y	—	P	L	I	T	7.8	0.78	+
S	F	N	Y	—	P	Q	V	T	9.1	0.85	+
S	F	N	Y	—	P	Q	D	T	9.4	0.87	+
S	F	N	Y	—	P	Q	I	I	8.3	0.52	+

[a] Octapeptides listed here are taken from Table 10 of Poorman *et al.* (1991). However, any duplicates, either to the octapeptides in the training set or to those in that table itself, are excluded. Therefore, the total testing octapeptides here should be 34 rather than 42.
[b] See footnote *a* to Table II.
[c] See footnote *b* to Table II.
[d] + or − represents cleavable or noncleavable by HIV-1 protease.
[e] Incorrect prediction by the correlation angle method.
[f] Incorrect prediction by the h probability method.

which is very close to the empirical rule for the normal distribution as described by Eq. (15). However, for the function h as used by Poorman *et al.* (1991), we instead have

$$\begin{cases} M \pm S & \text{contains } 56/105 \approx 53\% \\ & \text{of the predictions} \\ M \pm 2S & \text{contains } 105/101 = 100\% \\ & \text{of the predictions} \end{cases} \tag{17}$$

which completely violate the empirical rule of Eq. (15), meaning that the predicted results based on the h function method are significantly distorted from the normal distribution. This might be caused by the arbitrary assignment for parameters in the h function method, as discussed at the beginning of this paper. Especially when the training data are limited, the arbitrary assignment might affect the objective nature of the predicted results.

3.2. HIV-2 Protease

Again, for facilitating comparison, the training

set used in the h function method (Poorman *et al.*, 1992) for HIV-2 protease was used here. The set consists of 22 oligopeptides, with 21 octapeptides and one heptapeptide. Thus, similar to Eq. (14) for the case of HIV-1 protease, we also have the following equation for HIV-2 protease:

P_4	P_3	P_2	P_1	—	$P_{1'}$	$P_{2'}$	$P_{3'}$	$P_{4'}$
				⇓				
S	Q	N	Y	—	P	I	V	Q
E	E	E	L	—	A	E	C	F
T	Q	I	M	—	F	E	T	F[2]
G	Q	V	N	—	Y	E	E	F
G	G	N	Y	—	P	V	Q	H
A	E	E	L	—	A	E	I	F
P	F	A	A	—	A	Q	Q	R
R	Q	V	L	—	F	L	E	K
A	T	I	M	—	M	Q	R	G
S	L	N	L	—	P	V	A	K
	A	N	L	—	A	E	E	A
P	T	L	L	—	T	E	A	P
S	F	I	G	—	M	E	S	A
Y	E	E	F	—	V	Q	M	M
R	H	V	M	—	T	N	L	G
Y	I	S	A	—	A	E	L	R
G	L	A	A	—	P	Q	F	S
D	G	N	G	—	T	I	D	F
G	D	A	L	—	L	E	R	N
N	P	T	E	—	A	E	L	Q
R	Q	A	G	—	F	L	G	L

(18)

[2]Note that there is an error in the paper by Poorman *et al.* (1991); that is, this sequence was mistakenly counted as "TQIM-FETP."

where the arrow indicates the scissile bond by HIV-2 protease. According to Eq. (18), we can derived p_j^i(H-2) ($i = 1, 2, \ldots, 20$; $j = 4, 3, \ldots, 3', 4'$), which, together with the abundance of amino acids, are listed in Table VI.

Thus, for any given octapeptide, we can calculate its projection angle $\Theta_{H\text{-}2}(x)$ with the standard cleavable vector $\mathbf{\Phi}$(H-2) for HIV-2 according to Eqs. (6) and (10). The critical angle Θ_2^* for which we used the same optimization procedure as described above in finding the critical angle Θ_2^* for HIV-1 protease thus remains to be determined. Now the training data for the cleavable set were taken from the 22 oligopeptides in Table VI. But the training data for the noncleavable set were taken from the 139 octapeptides along the calmodulin sequence. This is because that, except GEVN-YEEF and YEEF-VQMM, no other HIV-2 cleavage sites whatsoever were detected for calmodulin even after it was completely denatured (Poorman *et al.*, 1991). Therefore, its $148 - 7 - 2 = 139$ octapeptides can serve as a benchmark for HIV-2 protease in checking overpredicted results. The process of optimal procedure is illustrated in Fig. 2, from which we find $\Theta_{H\text{-}2}^* = 76.2°$.

By setting such a value for $\Theta_{H\text{-}2}^*$ to predict the 139 noncleavable octapeptides taken from the sequence of calmodulin, 124 were found to have negative values of $\Delta\Theta_{H\text{-}2}(x)$, meaning the rate of correct prediction is 89% within the noncleavable training set. The predicted results for the 22 cleavable oligopeptides in the HIV-2 protease training set are listed in Table VII. There, the values of $\Delta\Theta_{H\text{-}2}(x)$ are all greater than zero,

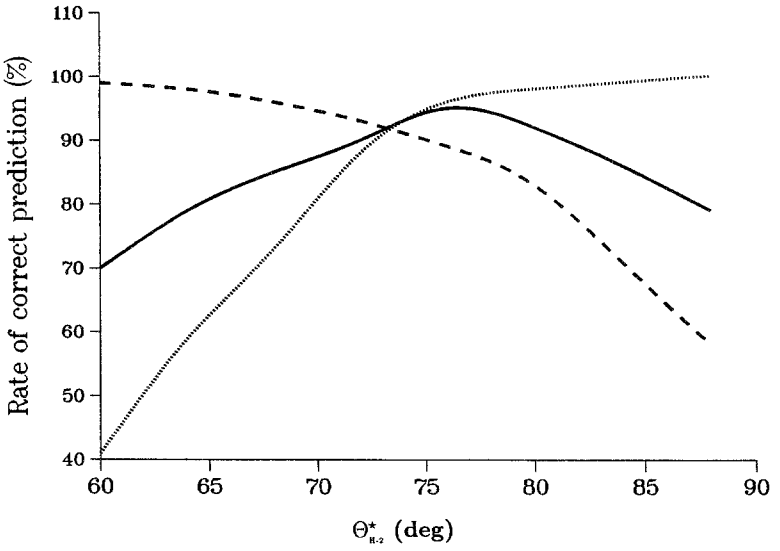

Fig. 2. Plot to show the dependence of the rate of corrected prediction on the critical angle $\Theta_{H\text{-}2}^*$: the rate of corrected prediction vs. $\Theta_{H\text{-}2}^*$ for the cleavable training set (i.e., the 22 oligopeptides listed in Table VII) is depicted by the dotted line; the rate of corrected prediction vs. $\Theta_{H\text{-}2}^*$ for the noncleavable training set (i.e., the 139 noncleavable octapeptides in calmodulin) is depicted by the dashed line; and an average of these two rates is depicted by the solid line. As shown by the plot, the larger the $\Theta_{H\text{-}2}^*$, the higher the rate of corrected prediction for the cleavable peptide set (see dotted line), but the lower the rate of corrected prediction for the noncleavable peptide set (see dashed line). As a consequence, their average reaches a maximum (see solid line) at $\Theta_{H\text{-}2}^* = 76.2°$, which is taken as the optimal parameter to predict the cleavability by HIV-2 protease according to Eq. (11).

Table IV. Predicted Results for Synthesized Analogs of HIV-1 Protease Substrate[a]

P_4	P_3	P_2	P_1	—	$P_{1'}$	$P_{2'}$	$P_{3'}$	$P_{4'}$	$\Delta\Theta_{H\text{-}1}(x)$(deg)[b]	h[c]	Hydrolysis[d]
S	Q	N	Y	—	P	I	V	Q	2.3	0.38	+
S	Q	N	Y	—	P	A	V	Q	0.7	0.20	+
S	Q	N	Y	—	P	N	V	Q	0.6	0.06[f]	+
S	Q	N	Y	—	P	F	V	Q	0.7	0.06[f]	+
S	Q	N	Y	—	P	L	V	Q	2.1	0.34	+
S	Q	N	Y	—	P	V	V	Q	2.4	0.38	+
S	Q	N	Y	—	P	G	V	Q	−0.3	0.03	−
S	Q	N	Y	—	P	D	V	Q	0.2[e]	0.05	−
S	Q	N	Y	—	P	K	V	Q	−0.1	0.04	−
S	Q	N	Y	—	A	I	V	Q	1.8[e]	0.28[f]	−
S	Q	N	Y	—	D	I	V	Q	−1.2	0.02	−
S	Q	N	Y	—	K	I	V	Q	−1.5	0.02	−

[a] Octapeptides listed here are taken from Bláha *et al.* (1991), where peptides with relative activity <0.01 are of resistance to HIV-1 protease (i.e. not cleaved).
[b] See footnote *a* to Table II.
[c] See footnote *b* to Table II.
[d] + and − refer to processing by, or resistance to, HIV-1 protease (Bláha *et al.*, 1991).
[e] Incorrect prediction by the projection angle method.
[f] Incorrect prediction by the *h* function method (Poorman *et al.*, 1991).

meaning that all the 22 oligopeptides can be cleaved by HIV-2 protease—namely, the method is fully self-consistent within the cleavable training set. However, using the *h* function method, it was found that the octapeptides NPTE-AELQ and RQAG-FLGQ were incorrectly predicted as noncleavable (Table VII).

The current method was further used to predict the cleavable sites in actin by HIV-2 protease. The actin sequence contains 377 residues and, hence, possesses 377 − 7 = 370 octapeptide sequences. Of these sequences, only TQIM-FETF, PTLL-TEAP, and SFIG-MESA are cleavable by HIV-2 protease (Poorman *et al.*, 1991). Thus, the remaining 367 octapeptides form a noncleavable set which is outside the above training set data and hence can serve as an independent set to test the predicted results. It has

Table V. Predicted Results for 15 Oligopeptides Cleavable by HIV-1 Protease as Observed Recently (Grifiths *et al.*, 1992)

P_4	P_3	P_2	P_1	—	$P_{1'}$	$P_{2'}$	$P_{3'}$	$P_{4'}$	$\Delta\Theta_{H\text{-}1}(x)$(deg)[a]	h[b]	Hydrolysis[c]
A	R	V	L	—	F	E	A	L	18.9	0.85	+
A	R	V	L	—	F	Q	A	L	10.1	0.74	+
A	R	V	L	—	F	I	A	L	7.8	0.51	+
A	R	V	L	—	F	V	A	L	8.0	0.51	+
A	R	V	L	—	F	A	A	L	6.3	0.29	+
A	R	V	L	—	F	D	A	L	5.8	0.07[d]	+
A	R	V	L	—	F	N	A	L	6.1	0.09[d]	+
A	R	V	L	—	F	T	A	L	6.4	0.24	+
A	R	N	L	—	F	E	A	L	17.6	0.86	+
A	R	V	Y	—	P	E	A	L	13.9	0.75	+
A	R	N	Y	—	P	E	A	L	12.6	0.76	+
S	Q	N	Y	—	P	I	V		2.5	0.49	+
S	Q	N	F	—	P	I	V	Q	7.8	0.67	+
S	Q	N	Y	—	P	I	V	L	2.4	0.46	+
A	Q	N	Y	—	P	I	V	L	3.2	0.48	+

[a] See footnote *a* to Table II.
[b] See footnote *b* to Table II.
[c] + Refers to processing by HIV-1 protease.
[d] Incorrect prediction by the *h* function method.

Table VI. Values of p_j^i Derived from Eq. 18 for Standard Vector Ψ (H-2)

Amino acid		Abundance[a]	Probability of amino acid i at each of 8 positions							
i	Code	μ^i	p_4^i	p_3^i	p_2^i	p_1^i	$p_{1'}^i$	$p_{2'}^i$	$p_{3'}^i$	$p_{4'}^i$
1	A	0.087	2/21	1/22	4/22	3/22	6/22	0/22	3/22	2/22
2	C	0.016	0/21	0/22	0/22	0/22	0/22	0/22	1/22	0/22
3	D	0.057	1/21	1/22	0/22	0/22	0/22	0/22	1/22	0/22
4	E	0.064	1/21	3/22	3/22	1/22	0/22	10/22	3/22	0/22
5	F	0.039	0/21	2/22	0/22	2/22	3/22	0/22	1/22	5/22[c]
6	G	0.078	4/21	2/22	0/22	3/22	0/22	0/22	1/22	2/22
7	H	0.022	0/21	1/22	0/22	0/22	0/22	0/22	0/22	1/22
8	I	0.052	0/21	1/22	3/22	0/22	0/22	2/22	1/22	0/22
9	K	0.068	0/21	0/22	0/22	0/22	0/22	0/22	0/22	2/22
10	L	0.082	0/21	2/22	1/22	7/22	1/22	2/22	3/22	1/22
11	M	0.021	0/21	0/22	0/22	3/22	2/22	0/22	1/22	1/22
12	N	0.044	1/21	0/22	6/22	1/22	0/22	1/22	0/22	1/22
13	P	0.045	3/21	1/22	0/22	0/22	5/22	0/22	0/22	1/22[c]
14	Q	0.039	0/21	5/22	0/22	0/22	0/22	4/22	2/22	3/22
15	R	0.048	3/21	1/22	0/22	0/22	0/22	0/22	2/22	2/22
16	S	0.066	3/21	0/22	1/22	0/22	0/22	0/22	1/22	1/22
17	T	0.058	1/21	2/22	1/22	0/22	3/22	0/22	1/22	0/22
18	V	0.070	0/21	0/22	3/22	0/22	1/22	3/22	1/22	0/22
19	W	0.012	0/21	0/22	0/22	0/22	0/22	0/22	0/22	0/22
20	Y	0.033	2/21	0/22	0/22	2/22	1/22	0/22	0/22	0/22

[a] The values of amino acid abundance in globular proteins are taken from Nakasgima *et al.* (1986).

Table VII. Predicted Results for 22 Oligipeptides in Training Set for HIV-2 Protease

P_4	P_3	P_2	P_1	—	$P_{1'}$	$P_{2'}$	$P_{3'}$	$P_{4'}$	$\Delta\Theta_{H\text{-}2}(x)$ (deg)[a]	h^b	Protein
S	Q	N	Y	—	P	I	V	Q	17.2	0.95	*gag*
E	E	E	L	—	A	E	C	F	31.2	0.95[c]	Troponin C
T	Q	I	M	—	F	E	T	F[c]	25.6	0.86[c]	Actin
G	Q	V	N	—	Y	E	E	F	25.2	0.93[c]	Calmodulin
G	C	N	Y	—	P	V	Q	H	12.9	0.90	*gag*
P	R	N	F	—	P	V	A	Q	14.2	0.90	*gag*
A	E	E	L	—	A	E	I	F	30.6	0.90[c]	Troponin C
P	F	A	A	—	A	Q	Q	R	12.1	0.86	*gag*
R	Q	V	L	—	F	L	E	K	14.9	0.85	*pol*
A	T	I	M	—	M	Q	R	G	4.3	0.84	*gag*
S	L	N	L	—	P	V	A	K	18.3	0.83	*pol*
	A	N	L	—	A	E	E	A	27.2	0.72	PE40
P	T	L	L	—	T	E	A	P	17.4	0.53[c]	Actin
S	F	I	G	—	M	E	S	A	12.0	0.67	Actin
Y	E	E	F	—	V	Q	M	M	1.2	0.63	Calmodulin
R	H	V	M		T	N	L	G	2.0	0.59	Calmodulin
Y	I	S	A	—	A	E	L	R	13.7	0.43	Calmodulin
G	L	A	A	—	P	Q	F	S	6.8	0.40	*pol*
D	G	N	G	—	T	I	D	F	7.1	0.45[c]	Calmodulin
G	D	A	L	—	L	E	R	N	16.6	0.33	PE40
N	P	T	E	—	A	E	L	Q	11.5	0.24[d]	Calmodulin
R	Q	A	G	—	F	L	G	L	3.0	0.24[d]	*gag*

[a] See Eq. (11), where $\Theta^*_{H\text{-}2} = 76.2°$ is derived through the optimization procedure as illustrated by Fig. 2.

[b] According to the h function method, a peptide can be cleaved by HIV-2 protease when its $h \geqslant h^* = 0.25$. Otherwise, it cannot be cleaved by the enzyme.

[c] Correction has been made here for the error found in the original Table 6 of Poorman *et al.* (1991) as indicated by the footnote in Eq. 18. This will affect the h values of those oligopeptides whose residues at the $P_{4'}$ subsite are P or F.

[d] Incorrect prediction by the h function method.

been found that, of the 367 noncleavable octapeptides, 350 have negative values of $\Delta\Theta_{H\text{-}2}(x)$, meaning the rate of correct prediction is 350/367 = 95%.

4. CONCLUSION

Octapeptides formed by 20 amino acids may form $20^8 = 2.56 \times 10^{10}$ different sequences. What kind of sequences can, and what kind of sequences cannot, be cleaved by HIV protease is a very important problem in designing effective HIV protease inhibitors as potential drugs for AIDS therapy. In view of this, a new method, the so-called projection angle method, is developed to predict the cleavability of a peptide sequence by HIV-1 or HIV-2 protease. The average predicted accuracy by the new method for the 105 peptide sequences whose cleavability is known to HIV-1 protease is 91.4%, which is about 8% higher than that of the existing *h* function method (Poorman *et al.*, 1991) for the same set of peptide sequences. A considerably high rate of correct prediction was also obtained when the new method was used to predict HIV-2 protease-cleaved sites in some proteins, although we still lack a sufficiently large cleavable training data set for HIV-2 protease. Moreover, the higher predicted rate is reflected by dealing with both the training set and testing set, indicating that the current method bears an improved feature in both self-consistency and extrapolating-effectiveness. Besides, the predicted results by the new method assume a normal distribution, but, the predicted results by the *h* function method do not, indicating that the new method is more reasonable than the previous one from the viewpoint of probability theory. It is expected that, with the accumulation of more experimental data on the cleavability of peptides by HIV protease, a better training set data can be established, and an even higher rate of prediction by the new method can be obtained. In view of this, the new method may be quite useful in the search for effective inhibitors of HIV protease, which is an important procedure in designing potential drugs for AIDS therapy.

REFERENCES

Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dauguet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W., and Montagnier, L. (1983). *Science* **220**, 868–871.

Bláha, I., Nemec, J., Tözsér, J., and Oroszlan, S. (1991). *Int. J. Peptide Protein Res.* **38**, 453–458.

Gallo, R. C., Salahuddin, S. Z., Popovic, M., Shearer, G. M., Kaplan, M., Haynes, B. F., Palker, T. J., Redfield, R., Oleske, J., Safai, B., White, G., Foster, P., and Markham, P. D. (1984). *Science* **224**, 500–503.

Griffiths, J. T., Phylip, L. H., Konvalinka, J., Strop, P., Gustchina, A., Wlpdawer, A., Davenport, R., Briggs, R., Dunn, B. M., and Kay, J. (1992). *Biochemistry* **31**, 5193–5200.

Hellen, C. U. T., Kräusslich, H. G., and Wimmer, E. (1989). *Biochemistry* **28**, 9881–9890.

Henderson, L. E., Benveniste, R. E., Sowder, R. C., Copeland, T. D., Schutz, A. M., and Oroszlan, S. (1988). *J. Virol.* **62**, 2587–2595.

Kohl, N. E., Emini, E. A., Schlief, W. A., Davis, L. J., Heimbach, J., Dixon, R. A. F., Scolnik, E. M., and Sigal, I. S. (1988). *Proc. Natl. Sci. USA* **85**, 4686–4690.

Mendenhall, W., Scheaffer, R. L., and Wackerly, D. D. (1986). *Mathematical Statistics with Applications*, Chap. 1, pp. 7–10.

Nakashima, H., Nishikawa, K., and Ooi, T. (1986). *J. Biochem.* **99**, 152–162.

Partin, K., Kräusslich, H. G., Ehrlich, L., Wimmer, E., and Carter, C. (1990). *J. Virol.* **64**, 3938–3947.

Poorman, R. A., Tomasselli, A. G., Heinrikson, R. L., and Kézdy, F. J. (1991). *J. Biol. Chem.* **266**, 14,554–14,561.

Putney, S. (1992). *TIBS* **17**, 191–196.

Schechter, I. and Borger, A. (1967) On the size of the active site in proteases. *J. Pain. Biochem. Biophys. Res. Com.* **27**, 157–162.

Wlodawer, A., Miller, M., Jaskólski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J., and Kent, S. B. H. (1989). *Science* **245**, 616–621.