

# A Formulation for Correlating Properties of Peptides and Its Application to Predicting Human Immunodeficiency Virus Protease-Cleavable Sites in Proteins

JAMES J. CHOU

Department of Physics, University of Michigan, Ann Arbor, Michigan 48104

## SYNOPSIS

A mathematical frame has been established to generally formulate the correlating properties of peptides. The formulation can be used to study the specificity of multisite enzymes, particularly in predicting the susceptible sites in proteins by human immunodeficiency virus (HIV) proteases, and hence can serve as a supplementary means in designing HIV protease inhibitors as potential drugs against acquired immunodeficiency syndrome. © 1993 John Wiley & Sons, Inc.

## INTRODUCTION

It has been clearly identified that human immunodeficiency virus (HIV) is the primary cause of acquired immunodeficiency syndrome (AIDS).<sup>1,2</sup> Therefore, a key step against AIDS is how to suppress HIV. It has been known that the replication of HIV is accompanied with 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.<sup>3-5</sup> 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.<sup>6</sup> In this regard, information about the HIV protease cleavage sites in polyproteins is very useful in refining our understanding of the specificity. And the knowledge thus acquired can play a guiding role for designing HIV protease inhibitors as potential drugs for AIDS therapy.<sup>4,7</sup> 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<sup>8</sup> 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  $P_1'$ . The model led to an algorithm to predict the cleavability by calculating its  $h$  value: if  $h \geq 0.13$ , the peptide sequence is able to be cleaved by HIV protease; otherwise, it is not. In calculating the probability  $h$ , the relevant parameters were derived from a set of peptide sequences whose cleavability by HIV protease is known. And the cutoff value 0.13 was set by a compromise between overpredicting and underpredicting. According to their report, the rate of correct prediction for 74 peptide sequences, of which 3 are heptapeptides and all the other octapeptides, are  $62/74 = 83.8\%$ . Considering the fact that the total number of octapeptides formed from 20 amino acids is  $(20)^8 = 2.56 \times 10^{10}$ , a predicted rate like that is encouraging. However, according to their method the value of  $h$  is a function of  $n_{i,j}$ , the frequency of the  $j$ th ( $j = 1, 2, \dots, 20$ ) amino acid occurring at the  $i$ th ( $i = 4, 3, 2, 1, 1', 2', 3', 4'$ ) subsite for a given training set of cleavable peptides. When  $n_{i,j}$  was zero, a value of 0.25 was arbitrarily assigned to it (see Ref. 8). This kind of arbitrary assignment would lead to the predicted results biased and subjective. In this paper, I would like to propose a new approach, the so-called correlation angle method, to predict



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

$$\left(\sum_{i=1}^n a_i b_i\right)^2 \leq \left(\sum_{k=1}^n a_k^2\right) \cdot \left(\sum_{k=1}^n b_k^2\right) \quad (8)$$

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

$$\begin{cases} \Theta_{H-1}(x) = \arccos \left\{ \frac{\Psi(x) \cdot \Psi(H-1)}{|\Psi(x)| |\Psi(H-1)|} \right\} \\ \Theta_{H-2}(x) = \arccos \left\{ \frac{\Psi(x) \cdot \Psi(H-2)}{|\Psi(x)| |\Psi(H-2)|} \right\} \end{cases} \quad (9)$$

or

$$\begin{cases} \Theta_{H-1}(x) = \arccos \left\{ \frac{\sum_{j=1}^4 \sum_{i=1}^{20} \psi_j^i(x) \psi_j^i(H-1)}{\left\{ \left[ \sum_{j=1}^4 \sum_{i=1}^{20} \psi_j^i(x)^2 \right] \left[ \sum_{j=1}^4 \sum_{i=1}^{20} \psi_j^i(H-1)^2 \right] \right\}^{1/2}} \right\} \\ \Theta_{H-2}(x) = \arccos \left\{ \frac{\sum_{j=1}^4 \sum_{i=1}^{20} \psi_j^i(x) \psi_j^i(H-2)}{\left\{ \left[ \sum_{j=1}^4 \sum_{i=1}^{20} \psi_j^i(x)^2 \right] \left[ \sum_{j=1}^4 \sum_{i=1}^{20} \psi_j^i(H-2)^2 \right] \right\}^{1/2}} \right\} \end{cases} \quad (10)$$

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

$$\begin{cases} \Delta\Theta_{H-1}(x) = \Theta_{H-1}^* - \Theta_{H-1}(x) \\ \Delta\Theta_{H-2}(x) = \Theta_{H-2}^* - \Theta_{H-2}(x) \end{cases} \quad (11)$$

where the parameters  $\Theta_{H-1}^*$  and  $\Theta_{H-2}^*$  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  $\Theta_{H-1}(x)$  or  $\Theta_{H-2}(x)$ , as can be formulated by the following equations:

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

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

The physical implication of Eqs. (12) and (13) can be further illustrated as follows. The vector  $\Psi(x)$  with  $\Delta\Theta_{H-1}(x) \geq 0$  [or  $\Delta\Theta_{H-2}(x) \geq 0$ ] has a greater projection on the standard vector  $\Psi(H-1)$  [or  $\Psi(H-2)$ ] than the vector  $\Psi(x)$  with  $\Delta\Theta_{H-1}(x) < 0$  [or  $\Delta\Theta_{H-2}(x) < 0$ ], and hence the peptide sequence corresponding to the former is more similar to the norm of the cleavable sequences than that of the latter. In other words, the current approach provides a quantitative description for the similarity of a peptide sequence  $x$  to the norm of the cleavable sequences through the correlation angle between their corresponding 160-D vectors.

It should be pointed out that this method, like the one proposed in Ref. 8, 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<sup>14</sup> and Ref. 8.

## 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  $\Theta_{H-1}(x)$  or  $\Theta_{H-2}(x)$ , the correlation angle of  $\Psi(x)$  with  $\Psi(H-1)$  or  $\Psi(H-2)$ , respectively. In order to realize that, we have to find  $\psi_j^i(H-1)$  ( $i = 1, 2, \dots, 20; j = 4, 3, \dots, 3', 4'$ ) or  $\psi_j^i(H-2)$  ( $i = 1, 2, \dots, 20; j = 4, 3, \dots, 3', 4'$ ), the 160 components for the standard vector  $\Psi(H-1)$  or  $\Psi(H-2)$ , respectively [see Eqs. (4) and (6)]. Because the abundance for each of the 20 amino acids in globular proteins is known,<sup>9</sup> if we can find  $p_j^i(H-1)$  or  $p_j^i(H-2)$ , we can immediately obtain  $\psi_j^i(H-1)$  or  $\psi_j^i(H-2)$  by means of Eq. (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 cleavable by HIV-1 or HIV-2 protease. Such a set of data is usu-

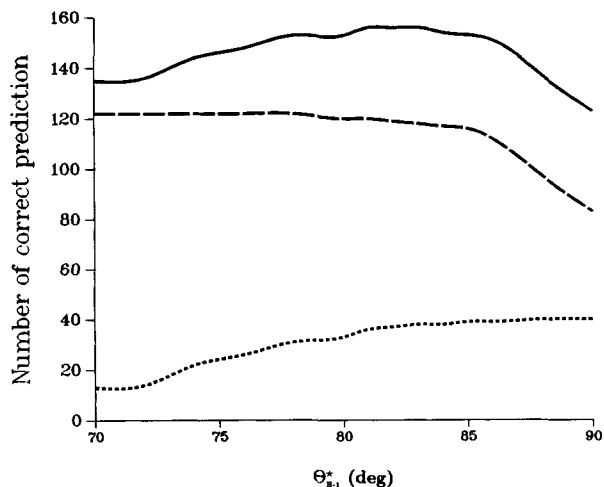
ally termed as "training set" or "development set." Below, let us first consider the case associated with HIV-1 protease. In order to compare the predicted results by means of the current correlation angle method with those by the  $h$  probability method, we should use the same set of training data. The following 40 peptide sequences, of which 38 are octapeptides and 2 heptapeptides, have been found at HIV-1 protease cleavage sites in proteins, i.e., can be cleaved by the enzyme:

P <sub>4</sub>	P <sub>3</sub>	P <sub>2</sub>	P <sub>1</sub>	—	P <sub>1'</sub>	P <sub>2'</sub>	P <sub>3'</sub>	P <sub>4'</sub>
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
S	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)

where the arrow indicates the scissile bond, which is between positions P<sub>1</sub> and P<sub>1'</sub>. The above 40 peptide sequences were used by the  $h$  probability method<sup>8</sup> as a training set for HIV-1 protease. Using the same training set, i.e., Eq. (14), we can derive  $p_j^i(H-1)$  ( $i = 1, 2, \dots, 20; j = 4, 3, \dots, 3', 4'$ ), which, together with the abundance of amino acids, are listed in Table I.

Based on Table I, for any given octapeptide we can calculate its correlation angle  $\Theta_{H-1}(x)$  with the standard cleavable vector  $\Psi(H-1)$  for HIV-1 according to Eqs. (5) and (10). Thus, according to Eq. (11), once the value of  $\Theta_{H-1}^*$  is determined,  $\Delta\Psi_{H-1}(x)$  is uniquely defined. To realize this, let us apply an optimization procedure, which is actually a 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  $\Theta_{H-1}^*$  is too large, then some noncleavable oligopeptides by the enzyme will be overpredicted as cleavable. On the other hand, if  $\Theta_{H-1}^*$  is too small, some cleavable oligopeptides will be underpredicted as noncleavable. Therefore, to find the optimal value for  $\Theta_{H-1}^*$ , two types of training



**Figure 1.** Plot to show the numbers of corrected prediction vs  $\Theta_{H-1}^*$ : the predicted results for the 40 cleavable oligopeptides listed in Table II are depicted by the dotted line, those for the 122 noncleavable octapeptides in the hen egg lysozyme by the dash line, and a combination of these two by the solid line. As shown by the plot, the larger the  $\Theta_{H-1}^*$ , the more the number of corrected prediction for the cleavable peptides (see dotted line), but the less the number of corrected prediction for the noncleavable peptides (see dash line). As a consequence, the total number of corrected prediction by adding these two would reach a maximum (see solid line) at  $\Theta_{H-1}^* = 82.4^\circ$ , which is taken as the optimal parameter for Eq. (11) to predict the cleavability of an oligopeptide by HIV-1 protease.

**Table I The Values of  $p_j^i$  Derived from Eq. (14) for the Standard Vector  $\Psi$  (H-1)**

Amino Acid		Abundance, <sup>a</sup> $\mu^i$	Probability of Amino Acid $i$ at Each of the Eight Positions							
Index	Code		$p^i$	$p_3^i$	$p_2^i$	$p_1^i$	$p_{1'}^i$	$p_{2'}^i$	$p_{3'}^i$	$p_{4'}^i$
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

<sup>a</sup> The values of amino acid abundance in globular proteins are taken from Ref. 9.

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 2. The training data for the latter can be obtained as follows. Since no HIV-1 cleavage sites whatsoever were detected for the hen egg lysozyme even after the protein (with 129 residues) was completely denatured,<sup>8</sup> the 129 - 7 = 122 octapeptides in the protein constitute a set of noncleavable peptides, which can be used as training data for checking overpredicted results. The process for finding the optimal  $\Psi_{H-1}^*$  is illustrated in Figure 1, where the number of correct prediction vs  $\Psi_{H-2}^*$  for the 40 cleavable oligopeptides is plotted by the dotted line, the number of correct prediction vs  $\Psi_{H-1}^*$  for the 122 noncleavable peptides is plotted by the dash line, and a combination by adding the above two is plotted by the solid line. As we can see from Figure 1, for the 40 cleavable oligopeptides, the larger the  $\Psi_{H-1}^*$ , the more the number of correct prediction; while for the 122 noncleavable octapeptides, the situation is just opposite. As a combined result of these two opposite changes with  $\Psi_{H-1}^*$ , there is a peak at  $\Psi_{H-1}^* = 82.4^\circ$  for the total number of correct prediction. Such a value is taken as the optimal value for  $\Theta_{H-1}^*$  because it leads to the highest rate of correct prediction for the training set data consisting of both cleavable and noncleavable oligopeptides.

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., RQANFLGK, ETTALVCD, and HLVEALYL, whose  $\Delta\Theta_{H-1}(x) < 0$ . This means that, except these three, all the other oligopeptides are correctly predicted as cleavable because the deviation of their characteristic vectors  $\Psi(x)$  from the standard cleavable vector for HIV-1 protease  $\Psi(H-1)$  is within the upper limit  $\Theta_{H-1}^*$ . However, according to the  $h$  probability method<sup>8</sup> in which the criterion for a cleavable peptide sequence is that its  $h$  value must be  $\geq 0.13$ , we find that there are eight incorrect prediction results. Consequently, the rate of correct prediction is 37/40 = 92.5% by using the current correlation angle method, while only 32/40 = 80.0% by the  $h$  probability method.<sup>8</sup>

The predicted results for the 34 octapeptides in a series of wild-type and mutant proteins<sup>10</sup> are listed in Table III. Note that although the octapeptides listed here are taken from Ref. 8, 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  $h$  probability method or the correlation 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

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

P <sub>4</sub>	P <sub>3</sub>	P <sub>2</sub>	P <sub>1</sub>	—	P <sub>1'</sub>	P <sub>2'</sub>	P <sub>3'</sub>	P <sub>4'</sub>	$\Delta\theta_{H-1}(x)$ , (deg) <sup>a</sup>	<i>h</i> <sup>b</sup>	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- $\beta$
S	F	N	F	—	P	Q	I	T	10.2	0.92	<i>pol</i>
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	<i>gag</i>
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	<i>gag</i>
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- $\beta$
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	<i>pol</i>
T	L	N	F	—	P	I	S	P	7.6	0.74	<i>pol</i>
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	<i>pol</i>
A	Q	T	F	—	Y	V	N	L	7.1	0.58	<i>pol</i>
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
S	A	N	L	—	A	E	E	A	13.2	0.39	PE40
S	Q	N	Y	—	P	I	V	Q	2.3	0.38	<i>gag</i>
P	G	N	F	—	L	Q	S	R	5.4	0.38	<i>gag</i>
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	<i>gag</i>
R	Q	A	N	—	F	L	G	K	-0.2 <sup>c</sup>	0.21	<i>gag</i>
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	<i>gag</i>
R	K	I	L	—	F	L	D	G	3.1	0.12 <sup>d</sup>	<i>pol</i>
G	S	H	L	—	V	E	A	L	9.0	0.10 <sup>d</sup>	Insulin
G	G	V	Y	—	A	T	R	S	0.9	0.10 <sup>d</sup>	Vimentin
F	R	S	G	—	V	E	T	T	5.6	0.09 <sup>d</sup>	<i>gag</i>
V	E	V	A	—	E	E	E	E	9.7	0.08 <sup>d</sup>	AAP
L	P	V	N	—	G	E	F	S	8.7	0.08 <sup>d</sup>	AAP
E	T	T	A	—	L	V	C	D	-4.8 <sup>c</sup>	0.03 <sup>d</sup>	Actin
H	L	V	E	—	A	L	Y	L	-1.8 <sup>c</sup>	0.02 <sup>d</sup>	Insulin

<sup>a</sup> See Eq. (11), where  $\theta_{H-1}^* = 82.4^\circ$  is derived through the optimization procedure as described in the text.

<sup>b</sup> According to the *h* probability method,<sup>8</sup> an octapeptide can be cleaved by HIV-1 protease when its *h*  $\geq 0.13$ . Otherwise, it cannot be cleaved by the enzyme.

<sup>c</sup> Incorrect prediction by the correlation angle method.

<sup>d</sup> Incorrect prediction by the *h* probability method.

than 42. These 34 oligopeptides are outside the training set and hence they 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,<sup>8</sup> the rate of

correct prediction for both methods is the same, equal to  $30/34 = 88.2\%$ .

Bláha et al.<sup>11</sup> have synthesized some analogues 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* probability

**Table III The Predicted Results for 34 Octapeptides<sup>a</sup> in a Series of Wild-Type and Mutant Proteins<sup>10</sup>**

P <sub>4</sub>	P <sub>3</sub>	P <sub>2</sub>	P <sub>1</sub>	—	P <sub>1'</sub>	P <sub>2'</sub>	P <sub>3'</sub>	P <sub>4'</sub>	$\Delta\theta_{H-1}(x)$ , (deg) <sup>b</sup>	<i>h</i> <sup>c</sup>	Experimental <sup>d</sup>
R	Q	N	Y	—	P	I	V	Q	1.4 <sup>e</sup>	0.34 <sup>f</sup>	—
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 <sup>e</sup>	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 <sup>e</sup>	0.02 <sup>f</sup>	+
S	K	N	Y	—	P	I	V	Q	-1.6 <sup>e</sup>	0.06 <sup>f</sup>	+
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 <sup>f</sup>	+
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	+

<sup>a</sup> Octapeptides listed here are taken from Table 10 of Ref. 8. 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.

<sup>b</sup> See footnote a to Table II.

<sup>c</sup> See footnote b to Table II.

<sup>d</sup> Plus or minus represents cleavable or noncleavable by HIV-1 protease, respectively.

<sup>e</sup> Incorrect prediction by the correlation angle method.

<sup>f</sup> Incorrect prediction by the *h* probability method.

method are listed in Table IV. As shown from the table, for 3 of the 12 oligopeptides, the predicted results are incorrect if the *h* probability 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 correlation angle method would

become apparently. According to the recent report by Griffiths et al.,<sup>12</sup> the 15 oligopeptides listed in Table V are definitely 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 2 of 15 events, the results were incorrectly predicted by the *h* function method, meaning the rate of correct prediction was 13/15 = 86.7%. But if using

**Table IV The Predicted Results for the Synthesized Analogues of HIV-1 Protease Substrate<sup>a</sup>**

P <sub>4</sub>	P <sub>3</sub>	P <sub>2</sub>	P <sub>1</sub>	—	P <sub>1'</sub>	P <sub>2'</sub>	P <sub>3'</sub>	P <sub>4'</sub>	$\Delta\theta_{H-1}(x)$ , (deg) <sup>b</sup>	<i>h</i> <sup>c</sup>	Hydrolysis <sup>d</sup>
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 <sup>f</sup>	+
S	Q	N	Y	—	P	F	V	Q	0.7	0.06 <sup>f</sup>	+
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 <sup>e</sup>	0.05	-
S	Q	N	Y	—	P	K	V	Q	-0.1	0.04	-
S	Q	N	Y	—	A	I	V	Q	1.8 <sup>e</sup>	0.28 <sup>f</sup>	-
S	Q	N	Y	—	D	I	V	Q	-1.2	0.02	-
S	Q	N	Y	—	K	I	V	Q	-1.5	0.02	-

<sup>a</sup> Octapeptides listed here are taken from Ref. 11, where peptides with relative activity < 0.01 are of resistance to HIV-1 protease, i.e., not cleaved.

<sup>b</sup> See footnote a to Table II.

<sup>c</sup> See footnote b to Table II.

<sup>d</sup> Plus and minus refer to processing by, or resistance to, HIV-1 protease,<sup>11</sup> respectively.

<sup>e</sup> Incorrect prediction by the correlation angle method.

<sup>f</sup> Incorrect prediction by the *h* probability method.<sup>8</sup>

the current sequence-coupled method, none of them was incorrectly predicted, i.e., a rate of correct prediction of 15/15 = 100%!

The average accuracy can be obtained by combining the data in Tables II–V, together, and it turns out to be 92/101 = 91.1% for the current method but only 84/101 = 83.2% for the *h* probability

method. This indicates that for the same set of data the average accuracy of the correlation angle method is about 8% higher than that of the *h* probability method.

According to statistical mathematics, a normal distribution should approximately satisfy the following empirical rule:<sup>13</sup>

**Table V The Predicted Results for 15 Oligopeptides Cleavable by HIV-1 Protease as Observed Recently<sup>12</sup>**

P <sub>4</sub>	P <sub>3</sub>	P <sub>2</sub>	P <sub>1</sub>	—	P <sub>1'</sub>	P <sub>2'</sub>	P <sub>3'</sub>	P <sub>4'</sub>	$\Delta\theta_{H-1}(x)$ , (deg) <sup>a</sup>	<i>h</i> <sup>b</sup>	Hydrolysis <sup>c</sup>
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 <sup>d</sup>	+
A	R	V	L	—	F	N	A	L	6.1	0.09 <sup>d</sup>	+
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	+

<sup>a</sup> See footnote a to Table II.

<sup>b</sup> See footnote b to Table II.

<sup>c</sup> Plus refers to processing by HIV-1 protease.

<sup>d</sup> Incorrect prediction by the *h* function method.



$$\left\{ \begin{array}{l} M \pm S \text{ contains approximately} \\ \qquad\qquad\qquad 66\% \text{ of the predictions} \\ M \pm 2S \text{ contains approximately} \\ \qquad\qquad\qquad 95\% \text{ of the predictions} \\ M \pm 3S \text{ contains almost all of the predictions} \end{array} \right. \quad (15)$$

where  $M$  and  $S$  represent mean and standard deviation of the predicted quantity, respectively. Now let us see what distribution we have for the 86 predicted results. Based on the data listed in Tables II-V, it is found that for  $\Delta\theta_{H-1}(x)$  we have

$$\left\{ \begin{array}{l} M \pm S \text{ contains } 67/101 \approx 66\% \text{ of predictions} \\ M \pm 2S \text{ contains } 96/101 \approx 95\% \text{ of predictions} \\ M \pm 3S \text{ contains } 101/101 = 100\%, \\ \qquad\qquad\qquad \text{i.e., all of predictions} \end{array} \right. \quad (16)$$

which is very close to the empirical rule for the normal distribution as described by Eq. (15). However, for the probability  $h$  as used in Ref. 8 we instead have

$$\left\{ \begin{array}{l} M \pm S \text{ contains } 54/101 \approx 53\% \text{ of predictions} \\ M \pm 2S \text{ contains } 101/101 = 100\% \text{ of predictions} \end{array} \right. \quad (17)$$

which completely violate the empirical rule of Eq. (15), meaning that the predicted results based on the  $h$  probability method are significantly distorted from the normal distribution. This might be caused by the arbitrary assignment for  $n_{i,j}$  in the  $h$  probability method, as discussed above. Especially when the training data are limited, the arbitrary assignment might affect the objective nature of the predicted results.

The current method can also be used to predict the cleavability of an octapeptide by the HIV-2 protease. In this case, however, the correlation angle should be calculated with respect to the standard vector  $\Psi(H-2)$  rather than  $\Psi(H-1)$ , and its components, as well as  $\theta_{H-1}^*$ , the upper limit of correlation angle for the oligopeptides cleavable by HIV-2 protease, should be derived instead from a training set for HIV-2 protease.

## 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 correlation 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 101 peptide sequences whose cleavability is known to HIV-1 protease is 91.1%, which is about 8% higher than that by the existing  $h$  probability method<sup>8</sup> for the same set of peptide sequences. 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  $h$  probability 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.

Finally, it is instructive to realize that, as pointed out by one of the referees of this paper, the new method is generally applicable to correlating properties of peptides beyond just their HIV protease cleavability.

## REFERENCES

1. Barré-Sinoussi, F., Chermann, J. C., Rey, F., Nugeyre, M. T., Chamaret, S., Gruest, J., Dautet, C., Axler-Blin, C., Vézinet-Brun, F., Rouzioux, C., Rozenbaum, W. & Montagnier, L. (1983) *Science* **220**, 868–871.
2. 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. & Markham, P. D. (1984) *Science* **224**, 500–503.
3. Kohl, N. E., Emimi, E. A., Schlieff, W. A., Davis, L. J., Heimbach, J., Dixon, R. A. F., Scolnik, E. M. & Sigal, I. S. (1988) *Proc. Natl. Sci. USA* **85**, 4686–4690.
4. Hellen, C. U. T., Kräusslich, H. G. & Wimmer, E. (1989) *Biochemistry* **28**, 9881–9890.

5. Wlodawer, A., Miller, M., Jaskólski, M., Sathyanarayana, B. K., Baldwin, E., Weber, I. T., Selk, L. M., Clawson, L., Schneider, J. & Kent, S. B. H. (1989) *Science* **245**, 616–621.
6. Putney, S. (1992) *TIBS*, **17**, 191–196.
7. Henderson, L. E., Benveniste, R. E., Sowder, R. C., Copeland, T. D., Schutz, A. M. & Oroszlan, S. (1988) *J. Virol.* **62**, 2587–2595.
8. Poorman, R. A., Tomasselli, A. G., Heinrikson, R. L. & Kézdy, F. J. (1991) *J. Biol. Chem.* **266**, 14554–14561.
9. Nakashima, H., Nishikawa, K. & Ooi, T. (1986) *J. Biochem.* **99**, 152–162.
10. Partin, K., Kräusslich, H. G., Ehrlich, L., Wimmer, E. & Carter, C. (1990) *J. Virol.* **64**, 3938–3947.
11. Bláha, I., Nemeč, J., Tózsér, J. & Oroszlan, S. (1991) *Int. J. Peptide Protein Res.* **38**, 453–458.
12. Griffiths, J. T., Phylip, L. H., Konvalinka, J., Strop, P., Gustchina, A., Wlodawer, A., Davenport, R., Briggs, R., Dunn, B. M. & Kay, J. (1992) *Biochemistry* **31**, 5193–5200.
13. Mendenhall, W., Scheaffer, R. L. & Wackerly, D. D. (1986) *Mathematical Statistics with Applications*, PWS-Kent, Boston, pp. 7–10.
14. Schechter, I. & Berger, A. (1967) *Biochem. Biophys. Res. Commun.* **27**, 157–162.

Received September 22, 1992

Accepted February 8, 1993