ANALYTICAL BIOCHEMISTRY **233,** 1–14 (1996)
Article No. 0001

REVIEW

Prediction of Human Immunodeficiency Virus Protease Cleavage Sites in Proteins

Kuo-Chen Chou

Computer-Aided Drug Discovery, Pharmacia & Upjohn Laboratories, Kalamazoo, Michigan 49007-4940

Knowledge of the polyprotein cleavage sites by HIV protease will refine our understanding of its specificity and the information thus acquired is useful for designing specific and efficient HIV protease inhibitors. The pace in searching for the proper inhibitors of HIV protease will be greatly expedited if one can find an accurate and rapid method for predicting the cleavage sites in proteins by HIV protease. Various prediction models or algorithms have been developed during the past 5 years. This Review is devoted to addressing the following problems: (1) Why is it important to predict the cleavability of a peptide by HIV protease? (2) What progresses have been made in developing the prediction methods, and what merits and weakness does each of these methods carry? The attention is focused on the state-of-the-art, which is featured by a discriminant function algorithm developed very recently as well as an improved database (the program and database are available upon request) established according to new experimental results. © 1996 Academic Press, Inc.

The human immunodeficiency virus (HIV), generally believed to be the causative agent (1, 2) of acquired immunodeficiency syndrome (AIDS), encodes an aspartic protease called the HIV protease whose function is essential for the replication of HIV (3–7). During the HIV life cycle, the precursor polyprotens are cleaved by the HIV protease. Loss of the cleavage ability results in the production of immature, noninfectious viral particles. Accordingly, HIV protease has been considered to be a promising target for the rational design of drugs against AIDS. As a complement to the strategy targeting another enzyme, the HIV reverse transcriptase (8), the design of HIV protease inhibitors represents a new approach to AIDS therapy (6, 9–12).

Functioning as a dimer, the HIV protease is made up of two identical subunits, each having 99 residues, but with only one active site. The active-site triad (Asp-25, Thr-26, Gly-27) is located in a loop whose structure is stabilized by a network of hydrogen bonds similar to that in the eukaryotic enzymes (13). The dimeric HIV protease has a crab-like shape (Fig. 1). A notable feature of the enzyme is that its catalytic cleft is "gated" by a pair of flaps (or pincers if viewed as a crab) formed each by a β hairpin of a monomer. Binding of an inhibitor (or a substrate) will induce a very large motion of the flap regions—as much as 7 Å for the ends of the flaps (12, 14). As a consequence, one has the following phenomenological picture: when the enzyme is in an inhibitor-free state, the flap-gate is open, allowing inhibitors or substrates to enter the binding cleft (Fig. 1a); when it is in an inhibitor-binding state, the flap-gate is closed, thereby blocking the entrance (Fig. 1b).

In order to inactivate HIV protease, knowledge about its specificity is particularly important. Many efforts have been made, trying to design HIV protease inhibitors by studying its substrate specificity (3, 5, 12, 14–16). As elucidated in the next section, the process of finding effective inhibitors will be greatly expedited if a rapid and accurate method is available to predict the cleavability of a peptide by HIV protease.

1. WHY IS IT IMPORTANT TO PREDICT THE CLEAVABILITY OF A PEPTIDE?

HIV protease is a member of the aspartyl proteases, a well-characterized mechanistic set of proteolytic enzymes in which the catalytic apparatus is made up of carboxyl groups derived from two aspartyl residues located in the N- and C-terminal halves of the enzyme molecule (17–21). These enzymes are highly substrate-selective and cleavage-specific, in that they cleave large, virus-specific polypeptides called polyproteins at defined amino acid pairs (5). It is known that the HIV protease-susceptible sites in a given protein extend to an octapeptide region (22), whose amino acid residues are sequentially symbolized by eight subsites R_4, R_3,

0003-2697/96 $12.00

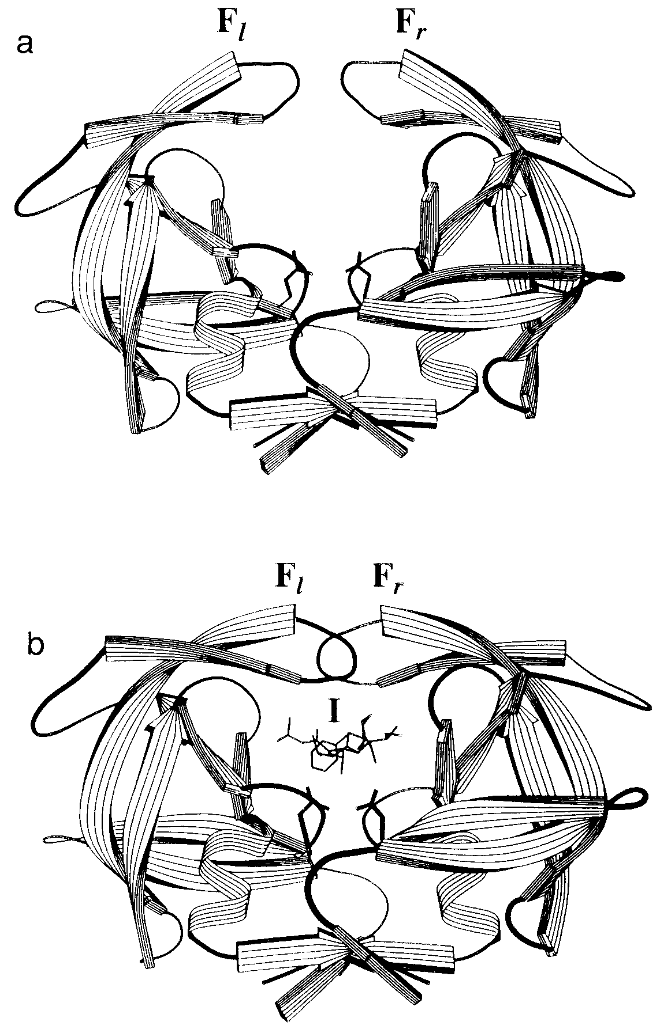

FIG. 1. Cartoon ribbon drawing of the dimer of HIV protease in (a) an inhibitor-free state (21), and (b) a complex with I, the inhibitor R0-31-8558 (7). F_l and F_r are the two flaps formed by the β hairpins of the left and right subunits, respectively, and they serve as a gate to control the entrance of inhibitors or substrates to the catalytic cleft. The whole molecule looks somewhat like a crab, with its pair of pincers likening to the two flaps. (Adapted, with permission, from Wlodawer and Erikson, Ref. 12.)

R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, $R_{4'}$ (Fig. 2). The reason here we use the symbol R rather than P as originally used by Schechter and Berger (23) is for avoiding any confusion with the symbol of the probability P introduced later. The scissile bond is located between the subsites R_1 and $R_{1'}$. Occasionally, the susceptible sites in some proteins may contain one subsite less or one subsite more (24, 25), corresponding to the case of an heptapeptide or nonapeptide, respectively. However, in studying the cleavability of peptide sequences by HIV proteases, heptapeptides and nonapeptides need to be considered only rarely.

Knowledge of the polyprotein cleavage sites by HIV protease will refine our understanding of its specificity and the information thus acquired is useful for designing specific and efficient HIV protease inhibitors. It is instructive to further elucidate this by the following rationale. According to the "lock-and-key" mechanism in enzymology, an HIV protease-cleavable peptide must satisfy the substrate specificity, i.e., a good fit for binding to the active site. (Here, the phrase of "good fit" should be understood in a broad sense rather than a narrow geometric sense; i.e., it means a favorable chemical-group disposition for the binding of a substrate to the active site of an enzyme and the catalytic reaction thereof.) However, such a peptide, after a modification of its scissile bond with some simple routine procedure, will completely lose its cleavability but it can still bind to the active site of an enzyme. Actually, the molecule thus modified can be compared with a "distorted key," which can be inserted into a lock but can neither open the lock nor be pulled out from it. That is why a molecule modified from a cleavable peptide can spontaneously become a competitive inhibitor against the enzyme. An illustration about such a concept is given in Fig. 2a, where an effective binding of a cleavable peptide to the active site of HIV protease is shown, while Fig. 2b shows that the peptide has become a noncleavable one after its scissile bond is modified, although it can still bind to the active site. Such a modified peptide, or "distorted key," will automatically

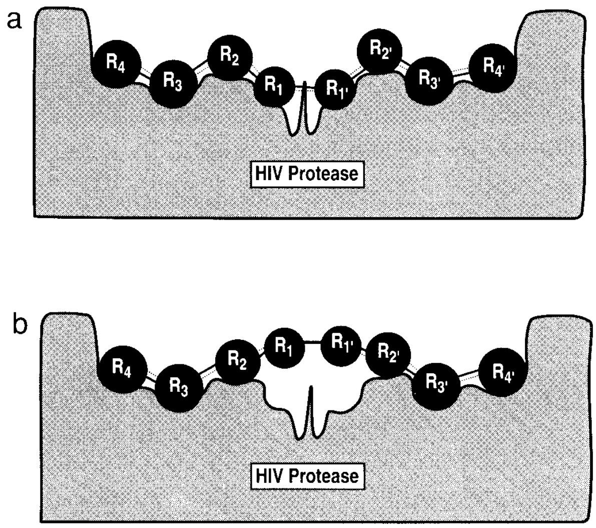

FIG. 2. Schematic illustration to show (a) a cleavable octapeptide is chemically effectively bound to the active site of HIV protease, and (b) although still bound to the active site, the peptide has lost its cleavability after its scissile bond is modified from a hybrid peptide bond (53) to a single bond by some simple routine procedure. The eight residues of the peptide are sequentially symbolized by R_4, R_3, R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, and $R_{4'}$. The scissile bond is located between R_1 and $R_{1'}$. The reason we use here the symbol R rather than P as introduced originally by Schechter and Berger (23) is to avoid confusion with the symbol of the probability P used later.

become an inhibitor candidate of HIV protease. Even for nonpeptide inhibitors, it also can provide useful insights about the key binding groups, proper microenvironment, and fitting conformation, as well as the requirement for hydrophobicity. Accordingly, in searching for the potential inhibitors, a matter of paramount importance is to discern what kind of peptides can be cleaved by HIV protease and what kind cannot be cleaved. Even limited in the range of an octapeptide, it is by no means easy to answer the question. This is because the number of possible octapeptides formed from 20 amino acids runs into $20^8 = 2.56 \times 10^{10}$. This is an astronomical figure! It would be exhausting to experimentally test out so many octapeptides. On the other hand, it would be very useful and would expedite our pace in search for the proper inhibitors of HIV protease if we could find an accurate and rapid method for predicting the cleavage sites in proteins by HIV protease. In view of this, various prediction methods have been developed during the past 5 years or so (25–31). This Review is devoted to discussing the progress of these methods, with a focus on the prospects in relation to the current "state-of-the-art."

Even for those who are sceptical about the importance of the specificity question in the design of drugs for HIV, it is still useful to present a systematic and comprehensive introduction for these methods due to their generality; i.e., they are applicable not only to HIV protease but also to any multisite enzymes.

2. TERM AND SYMBOL DEFINITIONS

For brevity and clarity, let us first give a unified definition for each of those terms or symbols that will repeatedly occur in various methods described in this Review. An octapeptide is generally expressed by

$$X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'},$$

where X_4 represents the amino acid at subsite R_4, X_3 represents the amino acid at subsite R_3, and so forth (Fig. 2). Two sets of peptides will be often mentioned in this Review: one is called cleavable or positive set, denoted by S^+, consisting of only cleavable peptides by HIV protease; and the other called noncleavable or negative set, S^-, consisting of only noncleavable peptides. Furthermore, we use S_1^+ or S_2^+ to represent the positive set composed of cleavable peptides by HIV-1 or HIV-2 protease, respectively. Likewise, we use S_1^- or S_2^- to represent the negative set composed of noncleavable peptides by HIV-1 or HIV-2 protease, respectively. The probability of amino acid X_i occurring at subsite R_i is expressed by $P(X_i)$: if it is derived from a cleavable set, the corresponding probability is expressed by $P^+(X_i)$; if derived from a noncleavable set, by $P^-(X_i)$. The conditional probability (32) that amino acid X_i occurring at the subsite position R_i given that X_j has occurred at position R_j is expressed by $P_i(X_i|X_j)$: if it is derived from a cleavable set, the corresponding conditional probability is expressed by $P_i^+(X_i|X_j)$; if derived from a noncleavable set, by $P_i^-(X_i|X_j)$.

3. PROGRESSES OF PREDICTION ALGORITHMS

Below, a brief description will be given for each of the existing prediction algorithms, as well as its merits and weakness.

3.1. The Cumulative Specificity Model or h-Function Algorithm

The cumulative specificity model was developed by Poorman *et al.* (25). The model postulates independent interactions of the eight amino acid moieties with their respective binding sites on the HIV protease. This model may be designated as the *h*-function method since the cleavability of a peptide is predicted according to the value of an *h* function that can actually be expressed by

$$h(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) = \frac{\prod_{i=4}^{4'} P_i^+(X_i)}{166 + \prod_{i=4}^{4'} P_i^+(X_i)}, \quad [1]$$

where $P_i^+(X_i)$ ($i = 4, 3, 2, 1, 1', 2', 3', 4'$) are defined in Section 2. However, in actual calculation, the values of $P_i^+(X_i)$ were modified by considering the abundance of each amino acid in globular proteins, and their values for HIV-1 and HIV-2 proteases are given in Tables 4 and 9 of Ref. (25). These data, symbolized as $s_{i,j}$ by them, were derived from 40 and 20 oligopeptides known to be cleavable by HIV-1 and HIV-2 proteases, respectively. A given peptide is predicted to be cleavable by HIV-1 or HIV-2 protease if the value of its respective *h* function is greater than $h_c = 0.13$ or 0.25. By means of this method, the rate of correct prediction for the 40 oligopeptides in the training set for HIV-1 protease was 32/40 = 80.0%, and that for a testing set of 34 octapeptides outside the training set was 30/34 = 88.2%. For such a complicated and intricate problem, a method with these predicted results should be deemed as promising. Moreover, since the only input for the algorithm to predict protease-susceptible sites in a given protein is its primary structure, it is suggested that the specificity of the enzyme is not directed toward any particular secondary structure but depends strongly on the accessibility of this segment. This finding has provided quite useful insights to the study of this field.

However, the *h*-function method suffers from the following three intrinsic shortcomings. (1) 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, not even the most neighboring coupling effect was taken into account along the peptide sequence. Obviously, this will certainly affect the accuracy of prediction. (2) As shown in Eq. [1], the *h* function was a multiplication of $P_i^+(\mathrm{X}_i)$. When, for a given training set S^+, the frequency of amino acid X_i occurring at subsite R_i is zero, an arbitrary value 0.5 had to be assigned for the corresponding modified $P_i^+(\mathrm{X}_i)$ (see Tables 4 and 9 of Ref. 25); otherwise, the *h* value would become zero no matter how favorable the specificity indices of the amino acids at the remaining subsites are. Such an arbitrary modification of $P_i^+(\mathrm{X}_i)$ might unduly influence the calculated results of *h*. (3) It should be noted that in the *h*-function method no clear procedure was described in determining the "cutoff value" h_c, a critical quantity in predicting the cleavability of an oligopeptide. The ambiguous treatment of such a critical quantity might introduce even more arbitrariness.

3.2. *The Vector Projection Model or Γ Function Algorithm*

In this algorithm (26), an octapeptide is expressed as a vector in an 8-D(dimensional) space, **V,** defined as:

$$\mathbf{V}(\mathrm{X}_4\mathrm{X}_3\mathrm{X}_2\mathrm{X}_1\mathrm{X}_{1'}\mathrm{X}_{2'}\mathrm{X}_{3'}\mathrm{X}_{4'}) = \begin{bmatrix} P_4^+(\mathrm{X}_4) - \tau(\mathrm{X}_4) \\ P_3^+(\mathrm{X}_3) - \tau(\mathrm{X}_3) \\ P_2^+(\mathrm{X}_2) - \tau(\mathrm{X}_2) \\ P_1^+(\mathrm{X}_1) - \tau(\mathrm{X}_1) \\ P_{1'}^+(\mathrm{X}_{1'}) - \tau(\mathrm{X}_{1'}) \\ P_{2'}^+(\mathrm{X}_{2'}) - \tau(\mathrm{X}_{2'}) \\ P_{3'}^+(\mathrm{X}_{3'}) - \tau(\mathrm{X}_{3'}) \\ P_{4'}^+(\mathrm{X}_{4'}) - \tau(\mathrm{X}_{4'}) \end{bmatrix}, \qquad [2]$$

where $P_i^+(\mathrm{X}_i)$, ($i = 4, 3, 2, 1, 1', 2', 3', 4'$), is defined in Section 2, and $\tau(\mathrm{X}_i)$ the mean abundance of amino acid X_i in globular proteins provided by Nakashima *et al.* (33). It is proper to use the Nakashima *et al.*'s database for the current purpose because it has been constructed from a database of reasonable size (3010 proteins) from which were eliminated the incomplete, short, and "closely related" sequences and also those proteins whose composition greatly deviates from the mean. By defining the components of the vector as the difference of these two frequencies as formulated by Eq. [2], the vector components represent the specificity of each subsite for the various amino acid residues. At the same time, the specificities are then normalized to zero: When the residue in a given subsite is indifferent, then the corresponding component is zero; the component is positive for amino acids toward which the subsite is specific. Conversely, a negative component reflects "negative specificity," i.e. an unfavorable influence of the substituent on the rate of the enzymatic reaction.

The cleavability of an octapeptide is calculated from the projection of its characteristic vector on an idealized, optimally cleavable vector. The larger the projection, the higher the likelihood that the peptide concerned can be cleaved by HIV protease. By introducing the approach of vector projection, the arbitrary value-assigning problem can be avoided even for the case of $P_i^+(\mathrm{X}_i) = 0$. This is an important advantage especially when the size of specificity database S^+ is very limited such as in the current case. Also, the cutoff value in this method was objectively determined via an optimization procedure between an overprediction and underprediction, which certainly represents an improvement compared with subjectively assigning a value for h_c as done in the *h*-function method. However, no sequence-coupled effect has been incorporated yet in this method.

3.3. *The Correlation-Angle Model or Θ Function Algorithm*

This is a very elegant algorithm, in which an octapeptide is expressed not by a vector in an 8-D space but one in an 8 × 20 = 160-D space (27). The bases of the 160-D space are actually a combination of the 8 subsites and the 20 native amino acids. The order of the former is from 4 to 4′ (Fig. 2), and that of the latter is numbered according to the alphabetic order of the single-letter amino acid code; i.e., $i = 1, 2, \ldots, 20$ for A (alanine), C (cysteine), . . . , Y (tyrosine), respectively. In the combination index, the array of 20 amino acids is counted first, followed by the array of 8 subsites. Thus, any octapeptide can be uniquely defined by a 160-D vector with either 1 or 0 as its components, depending on whether a base has a corresponding amino acid in the octapeptide concerned. For example, if an octapeptide is given by ACACYYYY, then its characteristic vector in the 160-D space is

$$\mathbf{\Psi}(\mathrm{ACACYYYY}) = \left\{\begin{matrix} 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, \\ 0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, \\ 1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, \\ 0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0, \\ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1, \\ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1, \\ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1, \\ 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1 \end{matrix}\right\} \qquad [3a]$$

Similarly, the norm of the cleavable set S^+ is defined by the 160-D vector

$$\mathbf{N}=\left\{\begin{matrix} n_4(\mathrm{A}),\ n_4(\mathrm{C}),\ n_4(\mathrm{D}),\ n_4(\mathrm{E}),\cdots,n_4(\mathrm{W}),\ n_4(\mathrm{Y}),\\ n_3(\mathrm{A}),\ n_3(\mathrm{C}),\ n_3(\mathrm{D}),\ n_3(\mathrm{E}),\cdots,n_3(\mathrm{W}),\ n_3(\mathrm{Y}),\\ n_2(\mathrm{A}),\ n_2(\mathrm{C}),\ n_2(\mathrm{D}),\ n_2(\mathrm{E}),\cdots,n_2(\mathrm{W}),\ n_2(\mathrm{Y}),\\ n_1(\mathrm{A}),\ n_1(\mathrm{C}),\ n_1(\mathrm{D}),\ n_1(\mathrm{E}),\cdots,n_1(\mathrm{W}),\ n_1(\mathrm{Y}),\\ n_{1'}(\mathrm{A}),n_{1'}(\mathrm{C}),n_{1'}(\mathrm{D}),n_{1'}(\mathrm{E}),\cdots,n_{1'}(\mathrm{W}),n_{1'}(\mathrm{Y}),\\ n_{2'}(\mathrm{A}),n_{2'}(\mathrm{C}),n_{2'}(\mathrm{D}),n_{2'}(\mathrm{E}),\cdots,n_{2'}(\mathrm{W}),n_{2'}(\mathrm{Y}),\\ n_{3'}(\mathrm{A}),n_{3'}(\mathrm{C}),n_{3'}(\mathrm{D}),n_{3'}(\mathrm{E}),\cdots,n_{3'}(\mathrm{W}),n_{3'}(\mathrm{Y}),\\ n_{4'}(\mathrm{A}),n_{4'}(\mathrm{C}),n_{4'}(\mathrm{D}),n_{4'}(\mathrm{E}),\cdots,n_{4'}(\mathrm{W}),\ n_{4'}(\mathrm{Y}) \end{matrix}\right\} \qquad [3b]$$

where $n_4(\mathrm{A}) = P_4^+(\mathrm{A}) - \tau(\mathrm{A})$, $n_3(\mathrm{C}) = P_3^+(\mathrm{C}) - \tau(\mathrm{C})$, and so forth. The probabilities $P_4^+(\mathrm{A})$, $P_3^+(\mathrm{C})$, $\cdots$ are defined in Section 2, and τ has the same definition as in Eq. [2]. The cleavability of an octapeptide is predicted according to the correlation angle defined by

$$\Theta = \arccos\left\{\frac{\mathbf{\Psi}\cdot\mathbf{N}}{|\mathbf{\Psi}||\mathbf{N}|}\right\} \qquad [3c]$$

The smaller the Θ, the higher the similarity between $\mathbf{\Psi}$ and **N,** and so is the likelihood that the peptide concerned can be cleaved by HIV protease. This method has all the merits as the vector projection method (26). Furthermore, there is no need to introduce an idealized, optimally cleavable vector, which is an additional merit compared with the vector projection method. However, the correlation angle method did not take sequence-coupled effect into account either.

3.4. *The Markov-Chain Model*

The Markov chain is a mathematical model in which the coupling effect is explicitly formulated through a conditional probability equation (34). According to this model, the criterion for predicting the cleavability of a given octapeptide $X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}$ is based on the function (28)

$$\begin{aligned}\Lambda(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) &= \log_{10}\{P_4^+(X_4)P_3^+(X_3|X_4)P_2^+(X_2|X_3)P_1^+(X_1|X_2)\\ &\quad\times \mathrm{P}_{1'}^+(X_{1'}|X_1)\ P_{2'}^+(X_{2'}|X_{1'})P_{3'}^+(X_{3'}|X_{2'})\\ &\quad\times P_{4'}^+(X_{4'}|X_{3'})\},\end{aligned} \qquad [4]$$

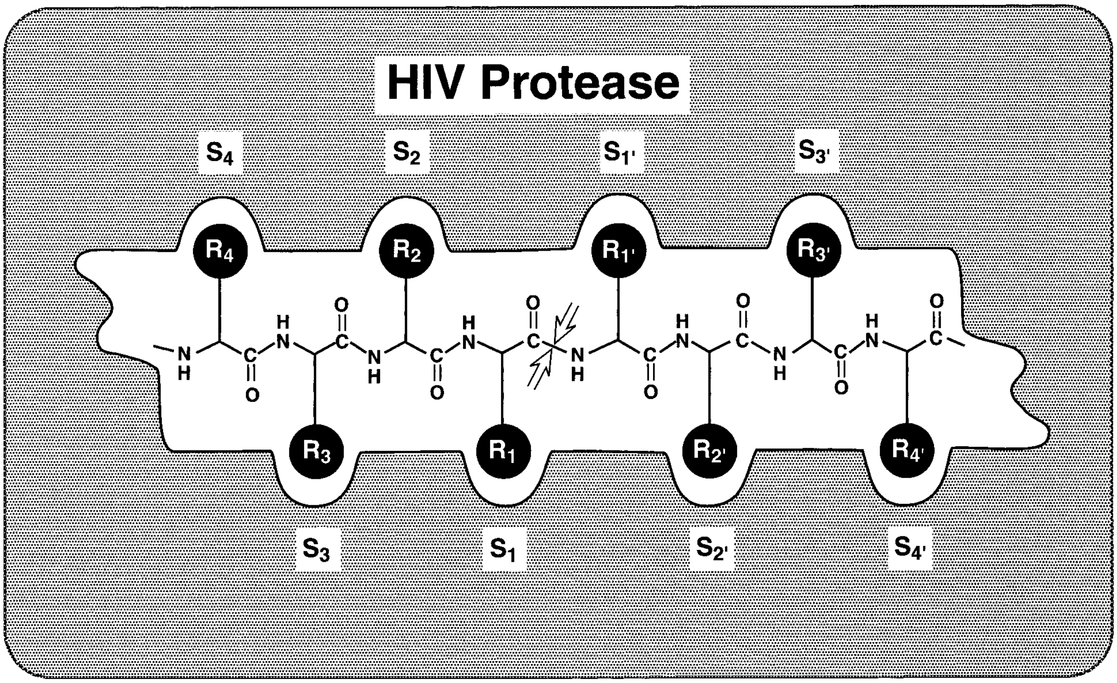

FIG. 3. Schematic representation of substrate bound to HIV protease based on analysis of protease-inhibitor crystal structures (12, 54, 55). The active site of enzyme is composed of eight extended "subsites," S_4, S_3, S_2, S_1, $S_{1'}$, $S_{2'}$, $S_{3'}$, and $S_{4'}$ and their counterparts in a substrate extend to an octapeptide region, sequentially symbolized by R_4, R_3, R_2, R_1, $R_{1'}$, $R_{2'}$, $R_{3'}$, and $R_{4'}$, respectively. The scissile bond is located between the subsites R_1 and $R_{1'}$. It can be seen that the side chains of the peptide substrate alternate between two completely opposite directions: R_4, R_2, $R_{1'}$, and $R_{3'}$ face one side, while R_3, R_1, $R_{2'}$, and $R_{4'}$ face the opposite side.

where the probability term $P_4^+(X_4)$ and the conditional probability terms $P_3^+(X_3|X_4)$, $P_2^+(X_2|X_3)$, and so forth are defined in Section 2. As expected, after incorporating the coupling effect among subsites, the rate of correct prediction was remarkably improved. However, in order to avoid the same situation that a zero value for any one of the factors in Eq. [4] would make the argument of logarithm abruptly become zero regardless of how large the other factors are, an arbitrary value was also assigned to replace the zero value for these factors as done in the *h*-function method (25).

3.5. *The Alternate-Subsite-Coupled Model*

In this model (29), the coupling effect among the subsites has been taken into account in a different manner. According to the protease-inhibitor crystal structures, the subsites face two completely opposite directions in an alternate way along its sequence (Fig. 3). Therefore, the side-chain interactions between residues i and $i + 2$ are stronger than those between i and $i + 1$ ($i = 1, 2, \ldots$) (35, 36). Furthermore, the "selectivity of subsite" (25) for R_2, R_4, and R_6 is higher than that for the remaining subsites, implying that the 2–4–6 correlation would play a dominant role in the sequence-coupling effect. To reflect such a coupling feature, instead of Eq. [4], the prediction algorithm should be based on the following formulation:

$$\begin{aligned}\prod(X_1X_2X_3X_4X_5X_6X_7X_8) &= \log_{10}\{P_1^+(X_1)P_2^+(X_2)P_3^+(X_3)P_4^+(X_4|X_2)P_5^+(X_5)\\ &\quad \times P_6^+(X_6|X_4)P_7^+(X_7)P_8^+(X_8)\}. \qquad [5]\end{aligned}$$

It can be seen from Eq. [5] that the 2–4–6 coupling is incorporated via the conditional probabilities $P_4^+(X_4|X_2)$ and $P_6^+(X_6|X_4)$. However, for the same reason as mentioned in the Markov-chain model, the arbitrary value-assigning problem cannot be avoided either.

3.6. *The Vectorized Sequence-Coupled Model*

This model is established based on two cornerstones: one is the sequence-coupled principle and the other is the vector-projection approach (30). By incorporating the sequence-coupled effect among the multiple subsites, the protease-cleavage mechanism can be more genuinely reflected; while by means of the vector-project approach, arbitrary assignment for insufficient experimental data can be avoided. Therefore, this model carries the merit of not only taking into account the coupling effect but also avoiding the arbitrary value-assigning problem. According to this model, an octapeptide is expressed by an 8-D vector formulated by

$$\mathbf{V}(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) = \begin{bmatrix} P_4^+(X_4)\\ P_3^+(X_3|X_4)\\ P_2^+(X_2|X_3)\\ P_1^+(X_1|X_2)\\ P_{1'}^+(X_{1'}|X_1)\\ P_{2'}^+(X_{2'}|X_{1'})\\ P_{3'}^+(X_{3'}|X_{2'})\\ P_{4'}^+(X_{4'}|X_{3'}) \end{bmatrix}, \qquad [6]$$

where the probability term $P_4^+(X_4)$ and the conditional probability terms $P_3^+(X_3|X_4)$, $P_2^+(X_2|X_3)$, and so forth are defined in Section 2. The cleavability of an octapeptide is predicted based on the projection of its characteristic vector $\mathbf{V}$ of Eq. [6] on an idealized, optimally cleavable vector. The larger the projection, the higher the likelihood that the peptide concerned can be cleaved by HIV protease. The threshold (or cutoff) value was determined via an optimization procedure between overprediction and underprediction.

4. THE DISCRIMINANT FUNCTION ALGORITHM

Much of this Review will be focused on this algorithm because it possesses all of the advantageous features carried by the previous algorithms. In addition, the tedious labor for deriving the cutoff value by the optimization procedure can be completely avoided because there is no need whatsoever to introduce such a quantity in the current algorithm.

According to the discriminant function algorithm (31), given an octapeptide, its attribute to the positive set S^+ or the negative set S^- can be formulated by an 8-D (dimension) vector $\mathbf{V}^+$ or $\mathbf{V}^-$, defined as

$$\mathbf{V}^+(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) = \begin{Bmatrix} P_4^+(X_4)\\ P_3^+(X_3|X_4)\\ P_2^+(X_2|X_3)\\ P_1^+(X_1|X_2)\\ P_{1'}^+(X_{1'}|X_1)\\ P_{2'}^+(X_{2'}|X_{1'})\\ P_{3'}^+(X_{3'}|X_{2'})\\ P_{4'}^+(X_{4'}|X_{3'}) \end{Bmatrix} \qquad [7a]$$

$$\mathbf{V}^-(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) = \begin{Bmatrix} P_4^-(X_4)\\ P_3^-(X_3|X_4)\\ P_2^-(X_2|X_3)\\ P_1^-(X_1|X_2)\\ P_{1'}^-(X_{1'}|X_1)\\ P_{2'}^-(X_{2'}|X_{1'})\\ P_{3'}^-(X_{3'}|X_{2'})\\ P_{4'}^-(X_{4'}|X_{3'}) \end{Bmatrix}, \qquad [7b]$$

where all the symbols have been defined in Section 2. Now in the 8-D space, let us define an ideal cleavability-positive vector, $\mathbf{\Lambda}^+$, 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. [7a]. Theoretically, the upper limit is 1, meaning that $\mathbf{\Lambda}^+$ would be the vector for a *hypothetical, idealized* oligopeptide which would be the only cleavable peptide for HIV protease. Therefore, for such an ideal cleavability-positive vector $\mathbf{\Lambda}^+$, 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 $\mathbf{\Lambda}^+$. The larger the projection, the higher the similarity, and hence the closer the peptide to the cleavability-positive set. Accordingly, the attribute function of a given octapeptide to the cleavability-positive set can be formulated by

$$\begin{aligned}\Psi^+(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) &= \mathbf{V}^+\cdot\mathbf{\Lambda}^+\\ &= P_4^+(X_4) + P_3^+(X_3|X_4) + P_2^+(X_2|X_3)\\ &\quad + P_1^+(X_1|X_2) + P_{1'}^+(X_{1'}|X_1)\\ &\quad + P_{2'}^+(X_{2'}|X_{1'}) + P_{3'}^+(X_{3'}|X_{2'}) + P_{4'}^+(X_{4'}|X_{3'}). \quad [8a]\end{aligned}$$

On the other hand, we can also in the 8-D space define an ideal cleavability-negative vector, $\mathbf{\Lambda}^-$, 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. [7b]. Theoretically, the upper limit is also 1, meaning that $\mathbf{\Lambda}^-$ would be the vector for a *hypothetical, idealized* oligopeptide 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

$$\begin{aligned}\Psi^-(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) &= \mathbf{V}^-\cdot\mathbf{\Lambda}^-\\ &= P_4^-(X_4) + P_3^-(X_3|X_4) + P_2^-(X_2|X_3)\\ &\quad + P_1^-(X_1|X_2) + P_{1'}^-(X_{1'}|X_1) + P_{2'}^-(X_{2'}|X_{1'})\\ &\quad + P_{3'}^-(X_{3'}|X_{2'}) + P_{4'}^-(X_{4'}|X_{3'}). \quad [8b]\end{aligned}$$

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. On the basis of this, let us define a discriminant function Δ given by

$$\begin{aligned}\Delta(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) &= \Psi^+(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'})\\ &\quad - \Psi^-(X_4X_3X_2X_1X_{1'}X_{2'}X_{3'}X_{4'}) + \Re \quad [9]\end{aligned}$$

where $\Re$ is a modified factor associated with some special empirical rules as will be described later (see Eq. [11]). If no empirical rules are incorporated, one may just set $\Re = 0$. Thus, the criterion for predicting the substrate specificity of a peptide can be formulated in terms of its discriminant function Δ as follows:

$$\begin{cases}\text{A peptide is cleavable by HIV protease,}\\ \qquad\qquad\qquad\qquad\text{if its } \Delta > 0\\ \text{A peptide is noncleavable by HIV protease,}\\ \qquad\qquad\qquad\qquad\text{otherwise.}\end{cases} \quad [10]$$

If, occasionally, the peptide to deal with is shorter than an octapeptide, such as a heptapeptide (25), we 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 Eq. [8] 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 should be S_1^+ and S_1^- which consist of the peptides associated with HIV-1 protease; while for the case of HIV-2 protease, the corresponding training set should be S_2^+ and S_2^- associated with HIV-2 protease.

It has been observed (37, 38) 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 $\Re$ for HIV-1 protease should be given as

$$\Re = \begin{cases}\Re_K, & \text{if K is at subsite } R_2, R_1, R_{1'} \text{ or } R_{2'},\\ 0, & \text{otherwise,}\end{cases} \quad [11]$$

where $\Re_K$ can be any large negative number as long as it can lead to $\Delta < 0$ (Eq. [9]) when the intolerable residue K occurs at any of the forbidden subsites. In this paper $\Re_K = -3$.

A question might be posed. As mentioned in Section 3.5 and illustrated in Fig. 3, for a substrate with an extended backbone conformation, the interactions between two alternative subsites along the sequence should be greater than those between two adjacent ones. Can the discrimination function algorithm as formulated by Eq. [8] reflect such a mechanism as characterized by a peptide with an extended backbone conformation? The answer is yes. This is because the current model is based on the Markov-chain theory (34), according to which the alternate-subsite-coupled effect is indirectly reflected. To make this clearer, let us give an illustration through the following simplified case. According to the current model (see Eq. [8]), the coupling effect for a segment of three amino acid sequence $X_{i-1}X_iX_{i+1}$ is given by

$$L(X_{i-1}X_iX_{i+1}) = P_i(X_i|X_{i-1}) + P_{i+1}(X_{i+1}|X_i). \quad [12]$$

On the other hand, according to the alternate-subsite-coupled model (29), the coupling effect for the same sequence should be expressed by (see Eq. [5])

$$L(X_{i-1}X_{i+1}) = P_i(X_{i+1}|X_{i-1}). \quad [13]$$

Since the normalization of conditional probabilities, it follows

$$\begin{aligned} L(X_{i-1}X_{i+1}) &= \sum_{\{Xi\}} L(X_{i-1}X_iX_{i+1}) \quad [14] \\ &= \sum_{\{Xi\}} \{P_i(X_i|X_{i-1}) + P_{i+1}(X_{i+1}|X_i)\}, \end{aligned}$$

where X_i represents any amino acid at subsite i and the summation is carried out over all the 20 amino acids. The above equation indicates that the alternate-subsite-coupled effect can be derived from the sequence-coupled effect, and hence the alternate-subsite-coupled model is only a special case of the sequence-coupled model. Accordingly, compared with the alternate-subsite-coupled model, the current model is more essential and general. It incorporates not only the coupling effect between subsites with adjacent positions but also that with alternative positions. In other words, more effects are taken into account in the current model than the alternate-subsite-coupled model. This is also reflected by the following fact. Compared with the alternate-subsite-coupled model (29), the sequence-coupled or Markov-chain model (28) is slightly better in predicting the cleavage sites in proteins by HIV protease although the results obtained by both methods were basically quite similar and consistent. Actually, the alternate-subsite-coupled model is an approximation of the sequence-coupled model, i.e., the case when the coupling effect between immediately adjacent subsites can be ignored.

5. DATABASE

In general any prediction method based on statistical theory is composed of two parts: one is the algorithm part, and the other is the database part. The last section is focused on the former, and this section will be focused on the latter.

In order to use any of the above algorithms, we need the data for $P^+(X_i)$, the probability of amino acid X_i occurring at subsite R_i. If the prediction is performed by the algorithm in which the coupling effect is incorporated, then we also need the data for $P_i^+(X_i|X_j)$, the conditional probability that amino acid X_i occurring at the subsite position R_i given that X_j has occurred at position R_j. Data of both these two types can be derived from a positive set S^+ consisting of cleavable peptides by HIV protease. In addition, if using any of the algorithms in Refs. (25–30), we also need the threshold (or cutoff) value, which can be determined by an optimization procedure between overprediction and underprediction. Or, if the prediction is performed by the discriminant function algorithm (31), then we shall instead need the data for $P^-(X_i)$ and $P_i^-(X_i|X_j)$. In either case, it requires a negative database S^- consisting of the noncleavable peptides only.

In the original paper by Poorman *et al.* (25) the training database S_1^+ for HIV-1 protease consists of 40 oligopeptides. However, according to some new experimental data reported recently (39), such a database should be extended to 62 oligopeptides, as given in Table 1. The training database S_1^- for HIV-1 protease consists of 239 noncleavable octapeptides (Appendix 1A),[1] 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 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-1 protease (25, 30). From these two tables we can derive $P^+(X_i)$, $P^-(X_i)$, $P_i^+(X_i|X_j)$, and $P_i^-(X_i|X_j)$, and their values are given in Appendices 1B, 1C, 1D, and 1E, respectively.

For HIV-2 protease, the positive training database S_2^+ consists of 22 oligopeptides (Table 2), where each is a substrate of HIV-2 protease (25). The negative training database S_2^- consists of 127 octapeptide (Appendix 2A), of which 122 are extracted from the sequence of hen egg lysozyme because no HIV-2 protease cleavage sites were ever detected even after it was completely denatured (30). And the other five octapeptides in S_2^- are derived from the octapeptide SQNYPIVQ in S_2^+ by substituting Pro at subsite $R_{1'}$ with Tyr, Phe, Leu, Met, and Val, respectively. This is not only because the peptides thus obtained are known not cleavable by HIV2 protease, but also because their hydrolysis by the enzyme are very sensitively dependent on the amino acid at $R_{1'}$ position (37). This kind of sensitivity cannot be reflected by the 122 octapeptides extracted from the hen egg lysozyme sequence alone. Therefore, the incorporation of the five additional octapeptides in S_2^- may, to some extent, reduce the case of overprediction. It should be realized that, owing to less experimental data reported for HIV-2 protease, the database for HIV-2 protease is relatively smaller, and hence the

[1] Owing to the space limit all the appendices mentioned in the text are not printed in the article. However, they are available from the author upon request.

TABLE 1

The Positive Training Database S_1^+ Consisting of 62 Cleavable Peptides by HIV-1 Protease[a]

Peptide sequence and cleavage site											
R_4	R_3	R_2	R_1	⇓	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	Δ[b]	$h - 0.13$[c]	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[d]
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[d]
L	P	V	N	⇓	G	E	F	S	2.69	−0.05	AAP[d]
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[e]
H	Y	G	F	⇓	P	T	Y	G	3.52	−0.13	NF-κB[f]
D	S	A	D	⇓	A	E	E	D	2.68	−0.11	AAP[d]
G	W	I	L	⇓	G	E	H	G	2.88	−0.08	LDH[g]
G	W	I	L	⇓	A	E	H	G	2.72	0.10	LDH

reliability is also lower. Nevertheless, with the improvement of database in both S_2^+ and S_2^-, the reliability of prediction by the discriminant function algorithm will certainly increase. The corresponding data derived from Table 2 and Appendix 2A for $P^+(X_i)$, $P^-(X_i)$, $P_i^+(X_i|X_j)$, and $P_i^-(X_i|X_j)$ are given in Appendices 2B, 2C, 2D, and 2E, respectively.

6. RESULTS AND DISCUSSION

As mentioned above, the discriminant function, or Δ-function, algorithm carries all the merits of the other algorithms, and hence it will be used here to demonstrate the predicted results. The predictions by the Δ-function algorithm have been performed for two sets of peptides, the training set and the testing set. The prediction for the former is a resubstitution examination to check the self-consistency of the new algorithm, while that for the latter is a cross-validation examination to check its extrapolating effectiveness. Below, let us examine the predicted results for the peptide cleavage sites by HIV-1 and HIV-2 proteases, respectively.

TABLE 1—*Continued*

Peptide sequence and cleavage site											
R_4	R_3	R_2	R_1	⇓	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	Δ[b]	$h - 0.13$[c]	Protein
Q	A	I	Y	⇓	L	A	L	Q	1.70	−0.13	*pol*[h]
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*[i]
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*[j]
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	K	E	E	H	G	3.03	0.05	Fibronectin[k]

[a] Note that listed here are 62 rather than 64 peptides as in Table 3 of Ref. (39) since two of them were chemically modified and should not be included here.

[b] 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 $\Delta \geqslant 0$; otherwise, it cannot be cleaved. The values of Δ were calculated according to Eqs. [8–9].

[c] h is the criterion used in the h-function method (25) to predict whether an octapeptide can be cleaved by HIV-1 protease: an oligopeptide can be cleaved when its $h \geqslant 0.13$; otherwise, it cannot be cleaved.

[d] Alzheimer amyloid protein.

[e] All entries to this point were referenced in Ref. (25).

[f] Riviere *et al.* (40).

[g] Tomaszek *et al.* (41).

[h] The following two entries are from Chattopadhyay *et al.* (42).

[i] The following seven entries are from Tomassellie *et al.* (39).

[j] The following eight entries are from Tözsér *et al.* (43).

[k] Oswald and von der Helm (44).

6.1. *HIV-1 Protease*

The Δ values calculated by Eq. [9] for the 62 oligopeptides in the cleavable set S_1^+ are given in Table 1, from which we can see that all have $\Delta > 0$, meaning that they are all correctly predicted to be cleavable by HIV-1 protease (see Eq. [10]). The Δ values calculated for the octapeptides in the noncleavable set S_1^- are given in Appendix 1A, from which we can see that, of the 239 noncleavable octapeptides, only 8 have $\Delta > 0$, i.e., are overpredicted to be cleavable. Therefore, the average rate of correct prediction for the training set data of S_1^+ and S_1^- is $(62 + 231)/(62 + 239) = 97.3\%$.

As a cross-validation, predictions have also been performed for a recently constructed testing set $\tilde{S}_1^+$ composed of 63 peptides which are known cleavable by HIV-1 protease but which are not included in the training set S_1^+. The predicted results are given in Appendix 1F, from which we can see that all but one are correctly predicted, and hence the prediction accuracy is $62/63 = 98.4\%$.

To provide an intuitive picture, a 3-D (dimension) histogram is depicted to show the predicted results for the peptides in the sets S_1^+, S_1^-, and $\tilde{S}_1^+$ (Fig. 4), where the peptides in each of the three sets are arranged from left to right along the abscissa according to their order in Table 1, Appendices 1A and 1F, respectively, and the corresponding Δ values are shown by the ordinate.

6.2. *HIV-2 Protease*

Similarly, the Δ values calculated by Eq. [9] for the 22 oligopeptides in the cleavable set S_2^+ are given in Table 2, from which we can see that all have $\Delta > 0$, meaning that they are all correctly predicted to be cleavable by HIV-2 protease. The Δ values calculated for the 127 octapeptides in the noncleavable set S_2^- are given in Appendix 2A, from which we can see that, of the 127 noncleavable peptides, only 4 are incorrectly predicted to be cleavable. Therefore, the average rate of correct prediction for the training set data of S_2^+ and S_2^- is $(22 + 123)/(22 + 127) = 97.3\%$.

As a cross-validation, predictions have also been performed for a testing set $\tilde{S}_2^+$ which consists of 51 peptides

TABLE 2

The Positive Training Database S_2^+ Consisting of 22 Cleavable Peptides by HIV-2 Protease[a]

Peptide sequence and cleavage site[b]										
R_4	R_3	R_2	R_1	⇓	$R_{1'}$	$R_{2'}$	$R_{3'}$	$R_{4'}$	Δ[b]	Protein
S	Q	N	Y	⇓	P	I	V	Q	1.17	*gag*
E	E	E	L	⇓	A	E	C	F	4.42	Troponin C
T	Q	I	M	⇓	F	E	T	F	2.88	Actin
G	Q	V	N	⇓	Y	E	E	F	3.34	Calmodulin
G	C	N	Y	⇓	P	V	Q	H	3.41	*gag*
P	R	N	F	⇓	P	V	A	Q	3.13	*gag*
A	E	E	L	⇓	A	E	I	F	3.88	Troponin C
P	F	A	A	⇓	A	Q	Q	R	2.54	*gag*
R	Q	V	L	⇓	F	L	E	K	3.11	*pol*
A	T	I	M	⇓	M	Q	R	G	2.96	*gag*
S	L	N	L	⇓	P	V	A	K	2.44	*pol*
	A	N	L	⇓	A	E	E	A	2.93	PE40
P	T	L	L	⇓	T	E	A	P	2.72	Actin
S	F	I	G	⇓	M	E	S	A	2.43	Actin
Y	E	E	F	⇓	V	Q	M	M	4.48	Calmodulin
R	H	V	M	⇓	T	N	L	G	3.26	Calmodulin
Y	I	S	A	⇓	A	E	L	R	3.75	Calmodulin
G	L	A	A	⇓	P	Q	F	S	2.64	*pol*
D	G	N	G	⇓	T	I	D	F	3.57	Calmodulin
G	D	A	L	⇓	L	E	R	N	2.81	PE40
N	P	T	E	⇓	A	E	L	Q	4.66	Calmodulin
R	Q	A	G	⇓	F	L	G	L	3.20	*gag*

Rate of correct prediction by resubstitution[c] = 22/22 = 100%

[a] The 22 peptides listed here are taken from Table 6 of Ref. (25) with a correction that the last letter of the 3rd peptide "P" has been replaced by "F" (see Ref. 30).
[b] See footnote *b* to Table 1.
[c] See Ref. 62.

which are known cleavable by HIV-2 protease but which are not included in the training set S_2^+. The predicted results are given in Appendix 2F, from which we can see that all are correctly predicted to be cleavable by HIV-2 protease. The rate of correct prediction is 51/51 = 100%. Note that the peptides listed in Appendix 2F are actually derived from the peptide SQNYPIVQ by single amino acid substitution at its different subsites, and they are all cleavable by HIV-2 protease, as observed by Tözsér *et al.* (37). These peptides have a common feature, i.e., they all have Pro at the $R_{1'}$ position. A question is naturally raised: what will happen if the Pro at $R_{1'}$ is substituted by some other amino acids? According to the reports by Bláha *et al.* (56) and Tözsér *et al.* (37), the following amino acids were introduced into $R_{1'}$ position for SQNYPIVQ: Ala, Asp, Lys, Tyr, Phe, Leu, Met, Gly, Val, Ile, Ser, and Trp. And they found that none of these peptides was hydrolyzed by HIV-2 protease. Of these noncleavable peptides, the following seven peptides occur neither in the positive training database S_2^+ nor in the negative training database S_2^-: SQNYAIVQ, SQNYDIVQ, SQNYKIVQ, SQNYGIVQ, SQNYIIVQ, SQNYSIVQ, and SQNYWIVQ. Therefore, they can serve as an additional set of independent data for a further cross-validation. The calculated results of Δ for these peptides are -0.03, -0.18, -0.23, -0.25, -0.03, -0.16, and -0.20, respectively, indicating that they all are noncleavable by HIV-2 protease, fully in consistent with the observations by Bláha *et al.* (56) and Tözsér *et al.* (37).

Also, to provide an intuitive picture, a 3-D histogram is given to show the predicted results for the peptides in S_2^+, S_2^-, and $\tilde{S}_2^+$ (Fig. 5), where the peptides in each of these sets are arranged from left to right along the abscissa according to their order in Table 2, Appendices 2A and 2F, respectively, and the corresponding Δ values are shown by the ordinate.

7. CAVEATS

It should be realized that using any of the above algorithms to identify potential sites of proteolysis in proteins may sometimes result in an inconsistency between theoretical prediction and experimental observation, especially for the case of overprediction, due to the following factors. (1) *Inaccessibility to the enzyme.* Some peptide sites in folded, native proteins may be perfectly susceptible to cleavage by HIV protease but cannot be observed because of being 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 the required accessibility of the protease to the predicted site. (2) *Unfavorable location in priority competition.* As mentioned at the beginning, the cleavage site by HIV protease usually requires eight amino acids in peptide substrates (24, 25). Thus, the maximum number of the predicted sites of cleavage within a given sequence of, say, a dozen amino acids, may be as many as 12 − 8 + 1 = 5. However, if one of the sites within a limited sequence region is highly favored over the others, cleavage at this site will result in fragments that are too short to serve as substrates, thereby removing the other predicted cleavages from the picture although it is not quite clear yet how much more favorable a cleavage point needs to be in order to prevent experimental observation of hydrolysis at nearby susceptible sites. Consequently, the inconsistency thus caused between theoretical prediction and experimental observation is merely a fake appearance, and should be termed as a pseudo-inconsistency. Accordingly, the corresponding overprediction should also be termed as a pseudo-overprediction.

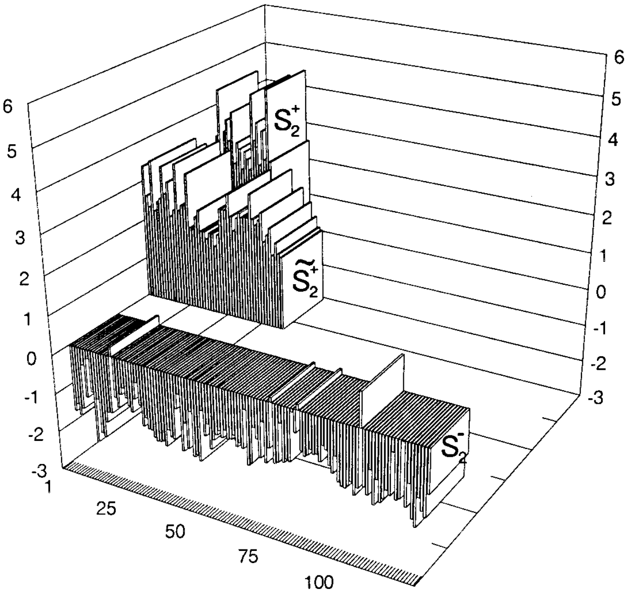

FIG. 5. The 3-D histogram to show the predicted Δ value for each of the 22 training cleavable peptides in S_2^+, the 122 training noncleavable peptides extracted from hen egg lysozyme in S_2^-, and the 51 testing cleavable peptides in $\tilde{S}_2^+$. The peptides in each of these three sets are arranged from left to right along the abscissa according to their order in Table 2, Appendices 2A, and 2F, respectively, and their Δ values are shown by the ordinate.

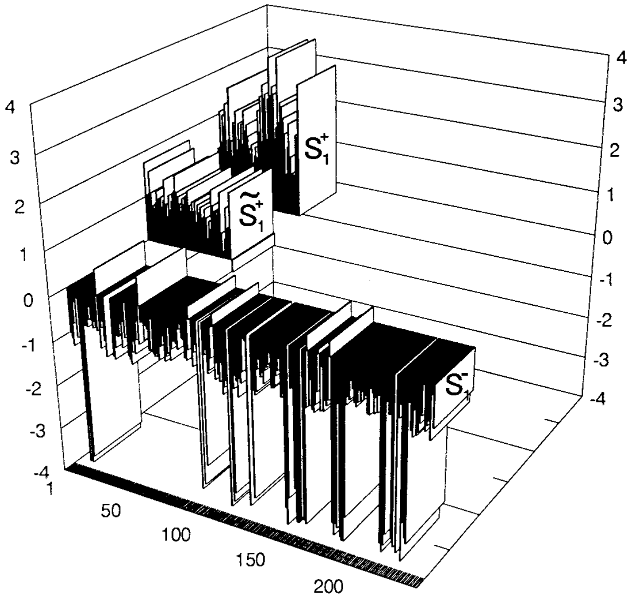

FIG. 4. The 3-D histogram to show the predicted Δ value for each of the 62 training cleavable peptides in S_1^+, the 239 training noncleavable peptides in S_1^-, and the 63 testing cleavable peptides in $\tilde{S}_1^+$. The peptides in each of these three sets are arranged from left to right along the abscissa according to their order in Table 1, Appendices 1A, and 1F, respectively, and their Δ values are shown by the ordinate.

8. CONCLUSIONS

The HIV protease cleavage sites in a protein are predictable from its primary structure. The accuracy of prediction can be significantly enhanced by incorporating the sequence-coupling effect into the prediction algorithm. It is equally important for improving the prediction accuracy by continuously updating the training database, of both positive and negative sets, based on newly accumulative experimental data. The vectorization approach makes any arbitrary value-assigning treatment unnecessary even for a very limited training database, and hence is an effectual measure to maintain the objectivity of prediction. The discriminant function algorithm developed recently not only carries all these advantageous features, such as sequence-coupled mechanism and vectorization approach, but also avoids the tedious labor for deriving the cutoff value before used to perform prediction.

Since 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 progresses of the relevant prediction algorithms will improve our ability and expedite the process for reaching this important therapeutic target.

It should be noted that the methods described here are general and can also be used to predict the substrate specificity of other multisite enzymes, such as GalNAc-transferase (57–59).

ACKNOWLEDGMENTS

The author thanks Dr. Tony E. Hugli for careful editing and critical reading of and commentary on the manuscript. Valuable discussions with Drs. F. J. Kézdy, G. M. Maggiora, C. T. Zhang, R. A. Poorman, A. G. Tomasselli, and R. L. Heinrikson are gratefully acknowledged. I also thank Dr. Douglas R. Morton, whose comments have stimulated a clear elucidation of how to use the predicted knowledge to help design potential drugs against AIDS. I am indebted to Raymond B. Moeller for drawing Figs. 2–5, and John Hui for providing his unpublished experimental results that neither HIV-1 cleavage sites in hen egg lysozyme and bovine pancreatic ribonuclease nor HIV-2 cleavage sites in hen egg lysozyme were ever detected even after they were completely denatured, as cited in the text. I particularly express special gratitude to Bi-Kun Luo and Prof. Hsiu-Chi Chou. Without their encouragement I would not have been able to overcome the difficulties that occurred during this study.

Note added in proof. During the stage of proof, the author has been notified that two artificial intelligence methods for predicting the cleavage sites in proteins by HIV protease are being developed and will be published soon. One is called the neural network algorithm (60), and the other is called the genetic algorithm (61).

REFERENCES

1. 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.
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., and Markham, P. D. (1984) *Science* **224,** 500–503.
3. 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. Acad. Sci. USA* **85,** 4686–4690.
4. Seelmeier, S., Schmidt, H., Turk, V., and von der Helm, K. (1988) *Proc. Natl. Acad. Sci. USA* **85,** 6612–6616.
5. Hellen, C. U. T., Kräusslich, H. G., and Wimmer, E. (1989) *Biochemistry* **28,** 9881–9890.
6. McQuade, T. J., Tomasselli, A. G., Liu, L., Karacostas, V., Moss, B., Sawyer, T. K., Heinrikson, H. L., and Tarpley, W. J. (1990) *Science* **247,** 454–456.
7. Graves, B. J., Hatada, M. H., Miller, J. K., Graves, M. C., Roy, S., Cook, C. M., Króhn, A., Martin, J. A., and Roberts, N. A. (1992) *in* Structure and Function of the Aspartic Protease: Genetics, Structure and Mechanisms (Dunn, B., Ed.), pp. 455–460. Plenum, New York.
8. Mitsuya, H., Yarchoan, R., and Broder, S. (1990) *Science* **249,** 1533–1544.
9. Meek, T. D., Lambert, D. M., Dreyer, G. B., Carr, J. T., Tomaszek, T. A., Jr., Moore, M. L., Strickler, J. E., Debouck, C., Hyland, L. J., Matthews, J. T., Metcalf, B. W., and Petteway, S. R. (1990) *Nature* **343,** 90–92.
10. Roberts, N. A., Martin, J. A., Kinchington, D., Broadhurst, A. V., Craig, J. C., Duncan, I. B., Galpin, S. A., Handa, B. K., Kay, J., Krohn, A., Lambert, R. W., Merrett, J. H., Mills, J. S., Parkes, K. E. B., Redshaw, S., Ritchie, A. J., Taylor, D. L., Thomas, G. J., and Machin, P. J. (1990) *Science* **248,** 358–361.
11. Ashorn, P., McQuade, T. J., Thaisrivongs, S., Tomasselli, A. G., Tarpley, W. J., and Moss, B. (1990) *Proc. Natl. Acad. Sci. USA* **87,** 7472–7476.
12. Wlodawer, A., and Erickson, J. W. (1993) *Annu. Rev. Biochem.* **62,** 543–585.
13. Davies, D. R. (1990) *Annu. Rev. Biophys. Biophys. Chem.* **19,** 189–215.
14. Tomasselli, A. G., Hui, J. O., Sawyer, T. K., Thaisrivongs, S., Hester, J. B., and Heinrikson, R. L. (1991) *in* Structure and Function of Aspartic Proteases (Dunn, B. M., Ed.), pp. 469–482, Plenum Press: New York.
15. Henderson, L. E., Benveniste, R. E., Sowder, R. C., Copeland, T. D., Schutz, A. M., and Oroszlan, S. (1988) *J. Virol.* **62,** 2587–2595.
16. Putney, S. (1992) *Trends in Biochem. Sci.* **17,** 191–196.
17. Toh, H., Ono, M., Saigo, K., and Miyata, T. (1985) *Nature* **315,** 691–691.
18. Pearl, L., and Taylor, W. (1987) *Nature* **329,** 351–354.
19. Lapatto, R., Blundell, T., Hemmings, A., Overington, J., Wilderspin, A., Wood, S., Merson, J. R., White, P. J., Danley, D. E., Geoghegan, K. F., Hawrylik, S. J., Lee, S. W., Scheld, K. G., and Hobart, P. M. (1989) *Nature* **342,** 299–302.
20. Navia, M. A., Fitzgerald, P. M. D., McKeever, B. M., Leu, C. T., Heimbach, J. C., Herber, W. K., Sigal, I. S., Darke, P. L., and Springer, J. P. (1989) *Nature* **337,** 615–620.
21. 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.
22. Miller, M., Schneider, J., Sathyanarayana, B. K., Toth, M. V., Marshall, G. R., L. M., Clawson, L., Selk, L., Kent, S. B. H., and Wlodawer, A. (1989) *Science* **246,** 1149–1152.
23. Schechter, I., and Berger, A. (1967) *Biochem. Biophys. Res. Commun.* **27,** 157–162.
24. 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., and Sigal, I. S. (1988) *Biochem. Biophys. Res. Commun.* **156,** 297–303.
25. Poorman, R. A., Tomasselli, A. G., Heinrikson, R. L., and Kézdy, F. J. (1991) *J. Biol. Chem.* **266,** 14554–14561.
26. Chou, K. C., Zhang, C. T., and Kézdy, F. J. (1993) *Proteins: Struct. Funct. Genet.* **16,** 195–204.
27. Chou, J. J. (1993) *J. Protein Chem.* **12,** 291–302.
28. Chou, K. C., and Zhang, C. T. (1993) *J. Protein Chem.* **12,** 709–724.
29. Zhang, C. T., and Chou, K. C. (1994) *Protein Eng.* **7,** 65–73.
30. Chou, K. C. (1993) *J. Biol. Chem.* **268,** 16938–16948.
31. Chou, K. C., Tomasselli, A. G., Reardon, I. M., and Heinrikson, R. L. (1996) *Proteins: Struct. Funct. Genet.* **24**(1), in press.
32. DeGroot, M. H. (1989) *in* Probability and Statistics, 2nd ed., Chap. 2, Addison–Wesley, Reading, MA.
33. Nakashima, H., Nishikawa, K., and Ooi, T. (1986) *J. Biochem.* **99,** 152–162.
34. Bhat, U. N. (1984) *in* Elements of Applied Stochastic Processes, Chap. 3, Wiley, New York.
35. Chou, K. C., Némethy, G., and Scheraga, H. A. (1983) *Biochemistry* **22,** 6213–6221.
36. Thanki, N., Rao, J. K., Foundling, S. I., Howe, W. J., Moon, J. B., Hui, J. O., Tomasselli, A. G., Heinrikson, R. L., Thaisrivongs, S., and Wlodawer, A. (1992) *Protein Sci.* **1,** 1061–1072.
37. Tözsér, J., Weber, I. T., Gustchina, A., Bláha, I., Copeland, T. D., Louis, J. M., and Oroszlan, S. (1992) *Biochemistry* **31,** 4793–4800.
38. Mildner, A. M., Rothrock, D. J., Leone, J., Bannow, C. A., Lull, J. M., Reardom, I. M., Sarcich, J. L., Howe, W. J., Tomich, C-S. C., Smith, C. W., Heinrikson, R. L., and Tomasselli, A. G. (1994) *Biochemistry* **33,** 9405–9413.
39. Tomasselli, A. G., Sarcich, J. L., Barrett, L. J., Reardon, I. M., Howe, W. J., Evans, D. B., Sharma, S. K., and Heinrikson, R. L. (1993) *Protein Sci.* **2,** 2167–2176.
40. Riviere, Y., Blank, V., Kourilsky, P., and Israel, A. (1991) *Nature* **350,** 625–626.

41. Tomaszek, T. A., Jr., Moore, M. L., Strickler, J. E., Sanchez, R. L., Dixson, J. S., Metcalf, B. W., Hassell, A., Dreyer, G. B., Brooks, I., Debouck, C., and Meek, T. D. (1992) *Biochemistry* **31,** 10153–10168.

42. 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., and Sharma, S. K. (1992) *J. Biol. Chem.* **267,** 14227–14232.

43. Tözsér, J., Blaha, I., Copeland, T. D., Wondrak, E. M., and Oroszlan, S. (1991) *FEBS Lett.* **281,** 77–80.

44. Oswald, M., and von der Helm, K. (1991) *FEBS Lett.* **292,** 298–300.

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

46. Bláha, I., Nemec, J., Tözsér, J., and Oroszlan, S. (1992) *Int. J. Peptide Protein Res.* **38,** 453–458.

47. 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.

48. Marczinovits, I., Molnár, J., and Patthy, A. (1994) *J. Biotechnol.* **37,** 79–83.

49. Freund, J., Kellner, R., Konvalinka, J., Wolber, V., Kräusslich, H. G., and Kalbitzer, H. R. (1994) *Eur. J. Biochem.* **223,** 589–593.

50. Hui, J. O., Tomasselli, A. G., Reardon, I. M., Lull, J. M., Brunner, D. P., Tomich, C-s C., and Heinrikson, R. L. (1993) *J. Protein Chem.* **12,** 323–327.

51. Tomasselli, A. G., Howe, W. J., Sawyer, T. K., Wlodawer, A., and Heinrikson, R. L. (1991) *Chimm. Oggi* **9,** 6–27.

52. 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., and Sharma, S. K. (1995) *J. Biol. Chem.* **270,** 13573–13579.

53. Schulz, G. E., and Schirmer, R. H. (1985) Principles of Protein Structure, Chap. 2, pp. 17–18, Springer-Verlag, New York.

54. Fitzgerald, P. M. D., and Springer, J. P. (1991) *Annu. Rev. Biophys. Chem.* **20,** 299–320.

55. Gustchina, A., Sansom, C., Prevost, M., Richelle, J., Wodak, S. Y., Wlodawer, A., and Weber, I. T. (1994) *Protein Eng.* **7,** 309–317.

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

57. Elhammer, Å. P., Poorman, R. A., Brown, E., Maggiora, L. L., Hoogerheide, J. G., and Kézdy, F. J. (1993) *J. Biol. Chem.* **268,** 10029–10038.

58. Chou, K. C., Zhang, C. T., Kézdy, F. J., and Poorman, R. A. (1995) *Proteins: Struct. Funct. Genet.* **21,** 118–126.

59. Chou, K. C. (1995) *Protein Sci.***4,** 1365–1383.

60. Thompson, T. B., Chou, K. C., and Zheng, C. (1995) *J. Theor. Biol.*, in press.

61. Noever, D., and Baskaran, S. (1995) *Biophys. J.*, in press.

62. Chou, K. C., and Zhang, C. T. (1995) *Crit. Rev. Biochem. Mol. Biol.* **30,** 275–349.