Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein

YU-DONG CAI,[1] XIAO-JUN LIU,[2] XUE-BIAO XU,[3] KUO-CHEN CHOU[4]

[1]*Shanghai Research Centre of Biotechnology, Chinese Academy of Sciences, Shanghai, 200233, People's Republic of China*

[2]*Institute of Cell, Animal and Population Biology, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, United Kingdom*

[3]*Department of Computing Science, University of Wales, College of Cardiff, Queens Buildings, Newport Road, P.O. Box 916, Cardiff CF2 3XF, United Kingdom*

[4]*Computer-Aided Drug Discovery, Upjohn Laboratories, Kalamazoo, Michigan 49001-4940*

Received 19 January 2001; Accepted 14 August 2001

Abstract: 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, robust, and rapid method for predicting the cleavage sites in proteins by HIV protease. In this article, a Support Vector Machine is applied to predict the cleavability of oligopeptides by proteases with multiple and extended specificity subsites. We selected HIV-1 protease as the subject of the study. Two hundred ninety-nine oligopeptides were chosen for the training set, while the other 63 oligopeptides were taken as a test set. Because of its high rate of self-consistency (299/299 = 100%), a good result in the jackknife test (286/299 = 95%) and correct prediction rate (55/63 = 87%), it is expected that the Support Vector Machine method can be referred to as a useful assistant technique for finding effective inhibitors of HIV protease, which is one of the targets in designing potential drugs against AIDS. The principle of the Support Vector Machine method can also be applied to analyzing the specificity of other multisubsite enzymes.

 J Comput Chem 23: 267–274, 2002

Key words: HIV Protease; Support Vector Machine; cleavage sites; self-consistency; Jackknife test

Introduction

HIV protease, encoded by human immunodeficiency virus (HIV), plays a very important role during the HIV life cycle. The mature and infectious viral particles can only be generated when the precursor polyproteins are cleaved by the HIV protease properly; otherwise, the viral particles are inactive.[1–5] Accordingly, HIV protease has been considered to be a promising target for the rational design of drugs against acquired immunodeficiency syndrome (AIDS). Actually, many effects have been made to understand the specificity of HIV protease and to design HIV protease inhibitors.[6–12]

It has been known that the HIV protease is a member of the aspartyl proteases, which are highly substrate-selective and cleavage-specific enzymes, and cleave polyproteins at defined amino acid pairs.[3] It also has been known that the HIV protease-susceptible sites in a given protein generally extend to an octapeptide region,[13] although occasionally to a heptapeptide or a nonapeptide region. To find out the potential competitive inhibitors against HIV protease, knowledge about the specificity of substrates, such as which peptides can and which peptides cannot be cleaved by the enzyme, is essential. However, the number of possible octapeptides formed from 20 amino acids is very large. It is time consuming and painful to test all of them by experiments. Therefore, an accurate, robust, and rapid method for predicting the cleavage sites of HIV protease would be very useful. During the last decade various approaches have been developed.[9–11, 14–18] For a comprehensive review, see Chou.[12] In this article, we apply Vapnik's Support Vector Machine[19] for predicting HIV protease cleavage sites in protein, and good results are obtained.

Support Vector Machine

The Support Vector Machine (SVM) is one kind of the learning machines based on statistical learning theory. The basic idea of applying SVM to pattern classification can be stated briefly as follows: first, map the input vectors into one feature space (possibly

Correspondence to: Y.-D. Cai; Biomolecular Sciences Department, UMIST, P.O. Box 88, Manchester, M60 1QD, UK; e-mail: y.cai@umist.ac.uk

DOI 10.1002/jcc.10017

with a higher dimension), either linearly or nonlinearly, which is relevant with the selection of the kernel function. Then, within the feature space from the first step, seek an optimized linear division, i.e., construct a hyperplane that separates two classes (this can be extended to multiclasses). SVM training always seeks a global optimized solution and avoids overfitting, so it has the ability to deal with a large number of features. A complete description of the theory of SVMs for pattern recognition is in Vapnik's book.[20]

SVMs have been used in a range of problems including drug design,[21] image recognition, and text classification.[22]

In this article, we apply Vapnik's Support Vector Machine[19] for predicting HIV protease cleavage sites in proteins. We downloaded the SVMlight, which is an implementation (in C Language) of SVM for the problem of pattern recognition. The optimization algorithm used in SVMlight can be found in Joachims.[23, 24] The code has been used in image recognition and text classification.[22]

Suppose we are given a set of samples, i.e, a series of input vectors

$$X_i \in R^d \qquad (i = 1, \ldots, N)$$

with corresponding labels $y_i \in \langle +1, -1 \rangle$ $(i = 1, \ldots, N)$. where -1 and $+1$ are used to stand, respectively, for the two classes. The goal here is to construct one binary classifier or derive one decision function from the available samples, which has small probability of misclassifying a future sample. Both the basic linear separable case and the most useful linear nonseparable case for most real-life problems are considered here.

The Linear Separable Case

In this case, there exists a separating hyperplane whose function is $\vec{W} \cdot \vec{X} + b = 0$, which implies:

$$y_i\left(\vec{W} \cdot \vec{x}_i + b\right) \geq 1, \qquad i = 1, \ldots, N$$

By minimizing $(1/2)\|\vec{W}\|^2$ subject to this constraint, the SVM approach tries to find a unique separating hyperplane. Here $\|\vec{W}\|^2$ is the Euclidean norm of $\vec{w}$, which maximizes the distance between the hyperplane (Optimal Separating Hyperplane or OSH in Cortes and Vapnik[25]) and the nearest data points of each class.The classifier is called the largest margin classifier.

By introducing Lagrange multipliers α_i, using the Karush–Kuhn–Tucker (KKT) conditions and the Wolfe dual theorem of the optimization theory, the SVM training procedure amounts to solving the following convex QP problem:

$$\text{Max:} \quad \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \cdot y_i y_j \cdot \vec{X}_i \cdot \vec{X}_j$$

subject to the following two conditions:

$$\alpha_i \geq 0 \qquad \sum_{i=1}^{N} \alpha_i y_i = 0, \qquad i = 1, \ldots, N$$

The solution is a unique globally optimized result can be shown having the following expansion:

$$\vec{W} = \sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x}_i$$

Only if the corresponding $\alpha_i > 0$, these $\vec{x}_i$ are called Support Vectors.

When a SVM is trained, the decision function can be written as:

$$f(\vec{x}) = \text{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot \vec{x} \cdot \vec{x}_i + b\right)$$

Where sgn( ) in the above formula is the given sign function.

The Linear Nonseparable Case

Two important techniques needed for this case are given respectively as below.

"Soft Margin" Technique

To allow for training errors, ref. 25 introduced slack variables:

$$\xi_i > 0, \qquad i = 1, \ldots, N$$

and relaxed separation constraint is given as:

$$y_i\left(\vec{w} \cdot \vec{X}_i + b\right) \geq 1 - \xi_i, \qquad (i = 1, \ldots, N)$$

The OSH can be found by minimizing

$$\frac{1}{2}\|\vec{W}\|^2 + C \sum_{i=1}^{N} \xi_i$$

instead of $1/2\|\vec{W}\|^2$ for the above two constraints in (see above), where C is a regularization parameter used to decide a trade-off between the training error and the margin.

"Kernel Substitution" Technique

SVM performs a nonlinear mapping of the input vector $\bar{x}$ from the input space R^d into a higher dimensional Hilbert space, where the mapping is determined by the kernel function. Then like in the case above, it finds the OSH in the space H corresponding to a nonlinear boundary in the input space.

Two typical kernel functions are listed below:

$$K\left(\vec{x}_i, \vec{x}_j\right) = \left(\vec{x}_i \cdot \vec{x}_j + 1\right)^d$$

$$K\left(\vec{x}_i, \vec{x}_j\right) = \exp\left(-r\left\|\vec{x}_i - \vec{x}_j\right\|^2\right)$$

where the first one is called the *polynomial kernel function of degree d*, which will eventually revert to the linear function when $d = 1$, the latter one is called the RBF (Radial Basic Function) kernel.

Finally, for the selected kernel function, the learning task amounts to solving the following QP problem,

$$\text{Max:} \quad \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j \cdot y_i y_j \cdot K\left(\vec{X}_i \cdot \vec{X}_j\right)$$

subject to:

$$0 \le \alpha_i \le C \qquad \sum_{i=1}^{N} \alpha_i y_i = 0, \qquad i = 1, \ldots, N$$

and the form of the decision function is

$$f(\vec{x}) = \operatorname{sgn}\left(\sum_{i=1}^{N} y_i \alpha_i \cdot K\left(\vec{x}, \vec{x}_i\right) + b\right)$$

For a given data set, only the kernel function and the regularity parameter C must be selected to specify one SVM.

The Training and Prediction of Cleavability

HIV proteases have extended substrate binding regions in which usually as many as eight consecutive amino acid moieties of the polypeptide substrate are in contact with the active site. We selected HIV-1 protease[12] as an object of study. In studying the specificity of HIV-1 protease, oligopeptides can be classified into two categories—the cleavable set and the noncleavable set. The cleavable set consists of oligopeptides, which can be cleaved by HIV-1 protease, while the noncleavable set consists of those that cannot be cleaved by HIV-1 protease.

In this research, the SVMlight,[23, 24] which is an implementation (in C Language) of SVM, was downloaded and applied to predict HIV protease cleavage sites in proteins.

In this research, 20 bases of oligopeptides are coded as 20-D vectors composed of only 0 and 1 (A = 100000...000, C = 010000...000, ..., Y = 000000...001), which are taken as the input to the SVM. The oligopeptides which can be cleaved by HIV-1 protease are taken as "positive" samples; the oligopeptides which can not be cleaved by HIV-1 protease are taken as "negative" samples.

Two hundren ninety-nine samples (from ref. 12), including 60 "positive" samples and 239 "negative" samples, are selected for the training set. In this research, for the SVM, the width of the Gaussian RBFs is selected as that which minimized an estimate of the VC-dimension (for detail, see refs. 23 and 24). The parameter C that controls the error-margin tradeoff is set at 5. The parameter γ is set at 1.1. This parameters selection of SVM is obtained from the comprision of the performance of the different parameters, which is detailed below. After being trained, the hyperplane output by the SVM was obtained. This indicates that the trained model, i.e., hyperplane output, which includes the important information, has the function to identify the cleavability of the peptide sequence.

Sixty-three samples (from ref. 12) including 54 "positive" samples and 9 "negative" samples are selected for the test set to be identified.

Table 1.

$\alpha(i)$	Support Vectors $X(i)$[a]							
0.361273	1:a	35:a	58:a	70:a	81:a	104:a	121:a	151:a
−1.307103	18:a	25:a	46:a	75:a	82:a	104:a	130:a	141:a
−0.802335	17:a	25:a	58:a	67:a	84:a	116:a	130:a	141:a
−2.033648	6:a	35:a	42:a	64:a	90:a	101:a	121:a	141:a
−2.775261	15:a	22:a	44:a	70:a	81:a	101:a	121:a	151:a
−0.098962	9:a	25:a	44:a	75:a	94:a	107:a	131:a	143:a
0.497676	1:a	35:a	58:a	70:a	85:a	114:a	121:a	150:a
0.160255	17:a	25:a	52:a	65:a	93:a	114:a	128:a	157:a
−0.183768	3:a	26:a	52:a	66:a	91:a	112:a	121:a	159:a
1.850334	17:a	34:a	43:a	65:a	99:a	104:a	138:a	154:a
−0.579963	1:a	21:a	49:a	65:a	84:a	116:a	132:a	145:a
−4.786837	1:a	21:a	41:a	69:a	85:a	104:a	135:a	154:a
−0.032423	17:a	34:a	41:a	72:a	89:a	107:a	128:a	148:a
4.439228	14:a	21:a	48:a	80:a	90:a	101:a	130:a	154:a
5.000000	4:a	37:a	57:a	61:a	90:a	118:a	122:a	143:a
4.074450	16:a	29:a	43:a	70:a	88:a	101:a	124:a	148:a
−1.870391	14:a	37:a	52:a	62:a	100:a	114:a	136:a	160:a
−1.850978	16:a	37:a	43:a	80:a	86:a	108:a	130:a	154:a
−0.624444	14:a	21:a	58:a	62:a	96:a	114:a	129:a	152:a
5.000000	15:a	34:a	41:a	72:a	85:a	110:a	126:a	149:a
2.931259	1:a	24:a	42:a	65:a	95:a	108:a	125:a	143:a
0.778213	3:a	37:a	58:a	70:a	84:a	104:a	131:a	156:a
−2.364755	13:a	40:a	58:a	73:a	98:a	107:a	125:a	143:a
−0.952818	10:a	34:a	48:a	72:a	96:a	115:a	139:a	159:a
−1.356589	16:a	34:a	49:a	72:a	98:a	101:a	122:a	149:a
1.950646	10:a	39:a	51:a	66:a	100:a	104:a	130:a	147:a
1.355379	4:a	30:a	44:a	70:a	81:a	104:a	132:a	155:a
3.480092	6:a	26:a	58:a	80:a	81:a	117:a	135:a	156:a
−2.422018	20:a	26:a	48:a	70:a	94:a	108:a	132:a	156:a
−2.817440	1:a	23:a	58:a	74:a	81:a	118:a	122:a	156:a
−1.306817	4:a	26:a	52:a	73:a	100:a	118:a	133:a	158:a
1.871299	16:a	25:a	48:a	66:a	91:a	104:a	136:a	141:a
2.992204	1:a	24:a	41:a	71:a	96:a	114:a	138:a	157:a
3.140189	20:a	24:a	44:a	65:a	98:a	114:a	131:a	151:a
−2.183628	2:a	24:a	50:a	61:a	81:a	101:a	131:a	149:a
−2.516052	2:a	24:a	46:a	72:a	93:a	120:a	138:a	153:a
−2.393721	8:a	28:a	58:a	61:a	82:a	104:a	126:a	152:a
2.394066	10:a	33:a	58:a	72:a	86:a	104:a	125:a	156:a
2.678836	13:a	35:a	52:a	65:a	93:a	118:a	121:a	154:a
5.000000	16:a	35:a	56:a	70:a	100:a	101:a	136:a	156:a
−0.369253	4:a	35:a	54:a	67:a	91:a	103:a	136:a	156:a
−2.709380	16:a	35:a	52:a	70:a	97:a	109:a	123:a	155:a
0.201815	6:a	39:a	48:a	70:a	86:a	104:a	127:a	146:a
0.846287	13:a	37:a	50:a	70:a	97:a	104:a	121:a	153:a
2.026203	18:a	24:a	48:a	62:a	97:a	104:a	131:a	144:a
3.342703	7:a	30:a	58:a	64:a	81:a	110:a	140:a	150:a
−1.728875	12:a	37:a	45:a	78:a	87:a	104:a	136:a	150:a
−2.983804	7:a	24:a	56:a	70:a	81:a	103:a	138:a	154:a
−1.238736	9:a	33:a	58:a	72:a	97:a	105:a	138:a	147:a

Results and Discussion

Success Rate of Self-Consistency and Prediction of SVMs

In this research, the examination for the self-consistency of the SVM method was tested. The rates of correct prediction for the two classes reaches 100%. This indicates that after being trained, the hyperplane output of the SVM has grasped the complicated relationship between the oligopeptides and cleavability, and it can be used to predict the unknown oligopeptides. The obtained support vectors $\vec{X}_i$ and α_i are shown in Table 1, and the threshold $b = 0.9538$ (For the detail of the decision function, see above).

Table 1. (Continued)

$\alpha(i)$	Support Vectors $X(i)$[a]							
3.307970	7:a	40:a	46:a	65:a	93:a	117:a	140:a	146:a
2.105288	16:a	30:a	52:a	70:a	93:a	118:a	121:a	149:a
2.236337	6:a	26:a	52:a	80:a	93:a	118:a	134:a	147:a
−2.692376	3:a	35:a	42:a	69:a	93:a	118:a	132:a	157:a
−1.188050	13:a	38:a	52:a	77:a	85:a	118:a	127:a	144:a
−2.097187	20:a	34:a	56:a	80:a	96:a	117:a	131:a	156:a
4.602633	6:a	23:a	41:a	80:a	85:a	116:a	138:a	153:a
3.235376	3:a	36:a	41:a	63:a	81:a	104:a	124:a	143:a
4.736785	13:a	25:a	41:a	61:a	81:a	114:a	134:a	155:a
−2.137380	4:a	30:a	41:a	61:a	81:a	111:a	129:a	155:a
−3.204175	16:a	30:a	41:a	63:a	98:a	114:a	121:a	158:a
−2.970450	2:a	36:a	41:a	70:a	90:a	116:a	136:a	143:a
3.270115	5:a	35:a	56:a	66:a	98:a	104:a	137:a	157:a
2.245763	1:a	34:a	57:a	65:a	100:a	118:a	132:a	150:a
−2.425965	17:a	23:a	60:a	66:a	88:a	110:a	134:a	148:a
−1.490272	6:a	36:a	57:a	63:a	100:a	106:a	128:a	150:a
−1.883787	6:a	32:a	53:a	80:a	98:a	113:a	138:a	147:a
2.831791	4:a	29:a	58:a	80:a	90:a	101:a	139:a	158:a
2.395776	14:a	28:a	57:a	70:a	99:a	114:a	135:a	153:a
2.819879	1:a	37:a	48:a	71:a	91:a	114:a	135:a	146:a
−3.062217	8:a	37:a	43:a	62:a	95:a	104:a	137:a	146:a
−2.141988	7:a	28:a	48:a	78:a	81:a	102:a	124:a	146:a
−1.452979	11:a	29:a	56:a	75:a	92:a	110:a	137:a	149:a
2.807846	6:a	36:a	47:a	70:a	98:a	104:a	121:a	150:a
2.149208	9:a	24:a	50:a	80:a	93:a	110:a	137:a	156:a
2.178159	10:a	24:a	58:a	72:a	88:a	118:a	137:a	143:a
−1.947573	13:a	38:a	47:a	65:a	83:a	101:a	136:a	158:a
−1.398801	5:a	24:a	56:a	72:a	85:a	112:a	137:a	154:a
−2.039602	10:a	23:a	52:a	80:a	95:a	106:a	140:a	156:a
1.817303	3:a	21:a	48:a	72:a	97:a	104:a	125:a	149:a
1.652432	15:a	29:a	48:a	70:a	85:a	110:a	123:a	146:a
2.001192	17:a	30:a	52:a	65:a	93:a	108:a	136:a	153:a
−1.913204	18:a	32:a	57:a	65:a	98:a	107:a	124:a	156:a
−1.250691	6:a	30:a	43:a	72:a	100:a	115:a	126:a	160:a
−1.505932	6:a	31:a	52:a	61:a	99:a	118:a	121:a	159:a
1.851407	13:a	26:a	52:a	65:a	90:a	114:a	136:a	155:a
−0.544877	15:a	27:a	46:a	70:a	83:a	112:a	140:a	155:a
−0.844327	15:a	34:a	47:a	71:a	83:a	116:a	136:a	157:a
−1.640160	16:a	35:a	52:a	70:a	82:a	112:a	128:a	153:a
1.432237	18:a	24:a	58:a	61:a	84:a	104:a	124:a	144:a
1.143071	1:a	24:a	57:a	65:a	100:a	117:a	123:a	146:a
−1.703916	1:a	39:a	48:a	75:a	86:a	102:a	135:a	150:a
−1.962638	4:a	36:a	52:a	65:a	92:a	117:a	134:a	141:a
−1.694050	13:a	22:a	56:a	61:a	90:a	110:a	136:a	156:a
2.501447	16:a	30:a	52:a	70:a	95:a	104:a	137:a	152:a
0.180938	6:a	23:a	41:a	70:a	90:a	104:a	135:a	152:a
−1.740200	1:a	29:a	45:a	64:a	95:a	114:a	127:a	151:a
−1.060533	20:a	36:a	57:a	71:a	96:a	108:a	137:a	143:a
−0.863256	14:a	21:a	57:a	72:a	95:a	112:a	137:a	143:a
−1.616218	4:a	37:a	41:a	61:a	81:a	109:a	125:a	144:a
−0.700769	12:a	29:a	47:a	68:a	88:a	118:a	121:a	142:a
−0.662418	18:a	33:a	58:a	67:a	85:a	103:a	121:a	156:a
−0.887580	3:a	26:a	55:a	77:a	93:a	106:a	136:a	155:a
−0.775868	12:a	40:a	55:a	66:a	100:a	116:a	130:a	146:a
−0.583747	16:a	32:a	45:a	72:a	97:a	114:a	121:a	157:a
1.096728	6:a	30:a	41:a	61:a	93:a	114:a	125:a	156:a
−1.080770	9:a	24:a	57:a	61:a	81:a	101:a	129:a	145:a
−0.626745	16:a	36:a	57:a	76:a	81:a	101:a	136:a	156:a
−0.753783	16:a	30:a	46:a	72:a	99:a	118:a	122:a	141:a

Table 1. (Continued)

$\alpha(i)$	Support Vectors $X(i)$[a]							
0.839884	13:a	28:a	58:a	66:a	81:a	104:a	137:a	145:a
1.233583	1:a	35:a	50:a	71:a	81:a	104:a	121:a	150:a
−1.174174	3:a	40:a	46:a	68:a	90:a	114:a	128:a	152:a
−0.414264	1:a	39:a	58:a	61:a	99:a	115:a	132:a	155:a
−1.090712	1:a	21:a	49:a	65:a	84:a	115:a	134:a	147:a
−1.041533	12:a	22:a	41:a	80:a	89:a	117:a	137:a	154:a
−0.834918	20:a	35:a	46:a	80:a	96:a	110:a	126:a	152:a
−0.768617	18:a	34:a	41:a	79:a	88:a	115:a	126:a	142:a
0.746678	1:a	24:a	44:a	70:a	81:a	104:a	128:a	145:a
−0.389604	18:a	27:a	44:a	76:a	90:a	101:a	123:a	158:a
−0.582874	16:a	21:a	50:a	70:a	96:a	116:a	123:a	148:a
−0.380949	10:a	21:a	41:a	61:a	91:a	109:a	135:a	147:a
−0.432350	9:a	26:a	57:a	63:a	98:a	114:a	121:a	159:a
−0.501881	17:a	21:a	41:a	61:a	89:a	105:a	124:a	155:a
−0.484537	6:a	34:a	57:a	72:a	82:a	120:a	134:a	156:a
−0.432558	16:a	36:a	49:a	80:a	93:a	112:a	122:a	141:a
0.373888	3:a	34:a	48:a	70:a	88:a	104:a	128:a	142:a
−0.376628	5:a	38:a	47:a	64:a	96:a	110:a	121:a	143:a
−0.542846	9:a	27:a	48:a	68:a	98:a	101:a	122:a	144:a
−0.336568	9:a	36:a	55:a	72:a	90:a	117:a	129:a	143:a
−0.456287	16:a	37:a	51:a	76:a	88:a	117:a	123:a	142:a
−0.410155	15:a	24:a	57:a	66:a	96:a	116:a	129:a	160:a
−0.414150	6:a	40:a	56:a	70:a	86:a	112:a	139:a	158:a
−0.220757	20:a	29:a	57:a	77:a	94:a	101:a	132:a	149:a
−0.255308	7:a	26:a	50:a	63:a	92:a	120:a	135:a	146:a
−0.202888	16:a	23:a	48:a	77:a	81:a	116:a	138:a	152:a
−0.336083	6:a	37:a	43:a	78:a	94:a	101:a	139:a	148:a
−0.194622	12:a	33:a	60:a	78:a	93:a	118:a	127:a	145:a
−0.210332	6:a	35:a	57:a	73:a	86:a	116:a	135:a	152:a
−0.155593	17:a	32:a	42:a	80:a	94:a	116:a	140:a	156:a
−0.202387	3:a	38:a	54:a	61:a	99:a	108:a	135:a	146:a
−0.184384	1:a	21:a	41:a	71:a	89:a	115:a	127:a	146:a
−0.164870	1:a	36:a	58:a	72:a	82:a	101:a	129:a	149:a
−0.061251	1:a	22:a	49:a	72:a	86:a	114:a	137:a	152:a
−0.038134	15:a	32:a	57:a	63:a	86:a	116:a	137:a	143:a
−0.018313	8:a	33:a	42:a	76:a	81:a	110:a	130:a	156:a
−0.013564	9:a	35:a	47:a	66:a	90:a	103:a	132:a	160:a
−0.007544	5:a	24:a	55:a	74:a	87:a	111:a	123:a	156:a

[a] a = 0.73105901, the others are 0.5.

Furthermore, to test the performance of the established model, 63 testing samples are recognized. As a result, the correct predicting rate reaches 55/63 = 87%. Here, there are eight samples: RQNYPIVQ (negative), SNNYPIVQ (positive), SQCYPIVQ (positive), SQNMPIVQ (positive), SQNYLIVQ (positive), SQNYPIIQ (positive), SQNYPNVQ (positive), SQNYTIVQ (positive) are incorrectly predicted (see Table 2). This is because the predicting rate is decided by the trained samples. The training set contain sample RCKGTDVQ (positive), which is similar to RQNYPIVQ, and sample SSKYPNCA (negative), which is similar to SNNYPIVQ, SQCYPIVQ, SQNMPIVQ, SQNYLIVQ, SQNYPIIQ, SQNYPNVQ and SQNYTIVQ.

Success Rate of Jackknife Test of SVMs

In statistical prediction, the jackknife test is deemed the most objective way for a crossvalidation test.[26–28] During the process of

Table 2. Incorrectly Predicted Samples.

Oligopeptides	Cleavability Predicted	Cleavability
RQNYPIVQ	positive	negative
SNNYPIVQ	negative	positive
SQCYPIVQ	negative	positive
SQNMPIVQ	negative	positive
SQNYLIVQ	negative	positive
SQNYPIIQ	negative	positive
SQNYPNVQ	negative	positive
SQNYTIVQ	negative	positive

jackknife analysis, both the training and testing datasets are actually open, and a peptide will, in turn, move from each to the other. The jackknife test results by the SVMs method indicated that the rate of correct prediction for the 299 samples is 286/299 = 95%.

Parameters Selection of SVMs

SVM's parameters includes kernel function, the regularization parameter C, γ for the RBFs, and degree d for the polynomial kernel function. In this research, we test the performance of the different parameters. The rates of self-consistency, prediction, and jackknife test for different parameters are shown in Table 3 (the linear kernel), Table 4 (the polynomial kernel), Table 5 (the RBFs kernel). From these tables, we can see that the selection of the kernel functions is most important; for this dataset, the RBFs is the best one. The regularization parameter C has some influence on the performance, for our dataset, we set C at 5. The parameters γ and d have little influence on the results, and we set γ at 1.1 for the dataset.

Comparison to Neural Network Method

We have applied the neural network method[29] to this problem. The comparison of the SVM method to the neural network is given in Table 6 (self-consistency test, jackknife test, and prediction).

Table 3. Performance of Linear Kernel.

Regularization Parameter	Rate of Correct Prediction (%)		
	Self-Consistency	Predicting for Unknown Smaples	Jackknife Test
0.5	84.62	19.05	81.94
1.0	92.64	34.92	89.63
1.5	95.65	57.14	92.64
2.0	97.32	80.95	92.98
2.5	97.66	85.71	93.65
3.0	97.66	85.71	93.31
3.5	97.99	82.54	92.98
4.0	98.66	82.54	92.31
4.5	99.00	82.54	91.64
5.0	99.00	84.13	91.97

Table 4. Performance of Polynomial Kernel.

Regularization Parameter C, Degree d	Rate of Correct Prediction (%)		
	Self-Consistency	Predicting for Unknown Smaples	Jackknife Test
(0.5,1.1)	84.62	19.05	81.94
(0.5,1.2)	84.62	19.05	81.94
(0.5,1.3)	84.62	19.05	81.94
(0.5,1.4)	84.62	19.05	81.94
(0.5,1.5)	84.62	19.05	81.94
(0.5,1.6)	84.62	19.05	81.94
(0.5,1.7)	84.62	19.05	81.94
(0.5,1.8)	84.62	19.05	81.94
(0.5,1.9)	84.62	19.05	81.94
(1.0,1.1)	92.64	34.92	89.63
(1.0,1.2)	92.64	34.92	89.63
(1.0,1.3)	92.64	34.92	89.63
(1.0,1.4)	92.64	34.92	89.63
(1.0,1.5)	92.64	34.92	89.63
(1.0,1.6)	92.64	34.92	89.63
(1.0,1.7)	92.64	34.92	89.63
(1.0,1.8)	92.64	34.92	89.63
(1.0,1.9)	92.64	34.92	89.63
(1.5,1.1)	95.65	57.14	92.64
(1.5,1.2)	95.65	57.14	92.64
(1.5,1.3)	95.65	57.14	92.64
(1.5,1.4)	95.65	57.14	92.64
(1.5,1.5)	95.65	57.14	92.64
(1.5,1.6)	95.65	57.14	92.64
(1.5,1.7)	95.65	57.14	92.64
(1.5,1.8)	95.65	57.14	92.64
(1.5,1.9)	95.65	57.14	92.64
(2.0,1.1)	97.32	80.95	92.98
(2.0,1.2)	97.32	80.95	92.98
(2.0,1.3)	97.32	80.95	92.98
(2.0,1.4)	97.32	80.95	92.98
(2.0,1.5)	97.32	80.95	92.98
(2.0,1.6)	97.32	80.95	92.98
(2.0,1.7)	97.32	80.95	92.98
(2.0,1.8)	97.32	80.95	92.98
(2.0,1.9)	97.32	80.95	92.98
(2.5,1.1)	97.66	85.71	93.65
(2.5,1.2)	97.66	85.71	93.65
(2.5,1.3)	97.66	85.71	93.65
(2.5,1.4)	97.66	85.71	93.65
(2.5,1.5)	97.66	85.71	93.65
(2.5,1.6)	97.66	85.71	93.65
(2.5,1.7)	97.66	85.71	93.65
(2.5,1.8)	97.66	85.71	93.65
(2.5,1.9)	97.66	85.71	93.65
(3.0,1.1)	97.66	85.71	93.31
(3.0,1.2)	97.66	85.71	93.31
(3.0,1.3)	97.66	85.71	93.31
(3.0,1.4)	97.66	85.71	93.31
(3.0,1.5)	97.66	85.71	93.31
(3.0,1.6)	97.66	85.71	93.31
(3.0,1.7)	97.66	85.71	93.31
(3.0,1.8)	97.66	85.71	93.31
(3.0,1.9)	97.66	85.71	93.31
(3.5,1.1)	97.99	82.54	92.98
(3.5,1.2)	97.99	82.54	92.98

Table 4. (Continued)

Regularization Parameter C, Degree d	Rate of Correct Prediction (%)		
	Self-Consistency	Predicting for Unknown Smaples	Jackknife Test
(3.5,1.3)	97.99	82.54	92.98
(3.5,1.4)	97.99	82.54	92.98
(3.5,1.5)	97.99	82.54	92.98
(3.5,1.6)	97.99	82.54	92.98
(3.5,1.7)	97.99	82.54	92.98
(3.5,1.8)	97.99	82.54	92.98
(3.5,1.9)	97.99	82.54	92.98
(4.0,1.1)	98.66	82.54	92.31
(4.0,1.2)	98.66	82.54	92.31
(4.0,1.3)	98.66	82.54	92.31
(4.0,1.4)	98.66	82.54	92.31
(4.0,1.5)	98.66	82.54	92.31
(4.0,1.6)	98.66	82.54	92.31
(4.0,1.7)	98.66	82.54	92.31
(4.0,1.8)	98.66	82.54	92.31
(4.0,1.9)	98.66	82.54	92.31
(4.5,1.1)	99.00	82.54	91.64
(4.5,1.2)	99.00	82.54	91.64
(4.5,1.3)	99.00	82.54	91.64
(4.5,1.4)	99.00	82.54	91.64
(4.5,1.5)	99.00	82.54	91.64
(4.5,1.6)	99.00	82.54	91.64
(4.5,1.7)	99.00	82.54	91.64
(4.5,1.8)	99.00	82.54	91.64
(4.5,1.9)	99.00	82.54	91.64
(5.0,1.1)	99.00	84.13	91.97
(5.0,1.2)	99.00	84.13	91.97
(5.0,1.3)	99.00	84.13	91.97
(5.0,1.4)	99.00	84.13	91.97
(5.0,1.5)	99.00	84.13	91.97
(5.0,1.6)	99.00	84.13	91.97
(5.0,1.7)	99.00	84.13	91.97
(5.0,1.8)	99.00	84.13	91.97
(5.0,1.9)	99.00	84.13	91.97

We can see the rate of the jackknife test of the SVM is better than that of the neural network, but its rate of prediction is worse, and its rate of self-consistency test is same as that of neural network.

Conclusion

Because of SVMs' strong ability in dealing with nonlinear problems such as predicting HIV protease cleavage sites in protein, it is quite reliable and accurate. So, it may be expected that the SVMs' method, the vectorized sequence-coupling method,[12] the discriminant function method,[18] as well as the other existing methods as described in a recent review,[12] if complemented with each other, will become an effective assistant tool in helping to finding effective inhibitors of HIV protease.

Table 5. Performance of RBFs.

Regularization Parameter C, γ	Rate of Correct Prediction (%)		
	Self-Consistency	Predicting for Unknown Smaples	Jackknife Test
(0.5,1.0)	86.62	20.63	82.61
(0.5,1.1)	87.96	20.63	82.94
(0.5,1.2)	87.96	20.63	83.61
(0.5,1.3)	88.29	20.63	83.95
(0.5,1.4)	88.96	20.63	84.28
(0.5,1.5)	89.30	20.63	84.28
(0.5,1.6)	90.30	20.63	84.28
(0.5,1.7)	90.30	20.63	84.28
(0.5,1.8)	90.30	20.63	84.28
(0.5,1.9)	90.30	20.63	84.28
(1.0,1.0)	95.32	46.03	89.30
(1.0,1.1)	95.65	47.62	89.63
(1.0,1.2)	96.32	53.97	89.63
(1.0,1.3)	97.32	53.97	89.63
(1.0,1.4)	97.66	53.97	90.64
(1.0,1.5)	97.66	55.56	90.97
(1.0,1.6)	97.66	55.56	90.97
(1.0,1.7)	97.99	55.56	90.97
(1.0,1.8)	97.99	57.14	90.97
(1.0,1.9)	98.33	58.73	90.97
(1.5,1.0)	97.66	73.02	92.31
(1.5,1.1)	98.33	76.19	92.31
(1.5,1.2)	98.66	76.19	92.31
(1.5,1.3)	99.33	76.19	92.31
(1.5,1.4)	99.33	76.19	92.31
(1.5,1.5)	99.33	76.19	92.31
(1.5,1.6)	99.33	76.19	92.64
(1.5,1.7)	99.67	76.19	92.64
(1.5,1.8)	100.00	76.19	92.31
(1.5,1.9)	100.00	76.19	92.31
(2.0,1.0)	99.33	79.37	93.31
(2.0,1.1)	99.33	79.37	93.31
(2.0,1.2)	99.33	79.37	93.98
(2.0,1.3)	99.33	79.37	93.98
(2.0,1.4)	100.00	80.95	93.98
(2.0,1.5)	100.00	80.95	93.65
(2.0,1.6)	100.00	79.37	93.65
(2.0,1.7)	100.00	77.78	93.65
(2.0,1.8)	100.00	76.19	93.31
(2.0,1.9)	100.00	76.19	93.31
(2.5,1.0)	99.33	82.54	93.98
(2.5,1.1)	100.00	82.54	93.98
(2.5,1.2)	100.00	84.13	93.98
(2.5,1.3)	100.00	84.13	93.98
(2.5,1.4)	100.00	84.13	93.98
(2.5,1.5)	100.00	82.54	93.98
(2.5,1.6)	100.00	82.54	93.98
(2.5,1.7)	100.00	80.95	93.65
(2.5,1.8)	100.00	80.95	93.31
(2.5,1.9)	100.00	79.37	93.31
(3.0,1.0)	100.00	87.30	93.98
(3.0,1.1)	100.00	87.30	94.31
(3.0,1.2)	100.00	87.30	94.31
(3.0,1.3)	100.00	85.71	93.98
(3.0,1.4)	100.00	84.13	93.98

Table 5. (Continued)

Regularization Parameter C, γ	Rate of Correct Prediction (%)		
	Self-Consistency	Predicting for Unknown Smaples	Jackknife Test
(3.0,1.5)	100.00	84.13	93.98
(3.0,1.6)	100.00	82.54	93.98
(3.0,1.7)	100.00	82.54	93.65
(3.0,1.8)	100.00	80.95	93.31
(3.0,1.9)	100.00	79.37	93.31
(3.5,1.0)	100.00	87.30	94.31
(3.5,1.1)	100.00	87.30	94.31
(3.5,1.2)	100.00	87.30	93.98
(3.5,1.3)	100.00	87.30	94.31
(3.5,1.4)	100.00	85.71	94.65
(3.5,1.5)	100.00	84.13	93.98
(3.5,1.6)	100.00	82.54	93.98
(3.5,1.7)	100.00	80.95	93.98
(3.5,1.8)	100.00	80.95	93.31
(3.5,1.9)	100.00	79.37	93.31
(4.0,1.0)	100.00	87.30	93.98
(4.0,1.1)	100.00	87.30	94.31
(4.0,1.2)	100.00	87.30	94.31
(4.0,1.3)	100.00	87.30	94.31
(4.0,1.4)	100.00	87.30	94.31
(4.0,1.5)	100.00	82.54	93.65
(4.0,1.6)	100.00	82.54	93.65
(4.0,1.7)	100.00	80.95	93.98
(4.0,1.8)	100.00	80.95	93.31
(4.0,1.9)	100.00	79.37	93.31
(4.5,1.0)	100.00	87.30	93.98
(4.5,1.1)	100.00	87.30	94.31
(4.5,1.2)	100.00	87.30	94.31
(4.5,1.3)	100.00	87.30	94.31
(4.5,1.4)	100.00	87.30	93.98
(4.5,1.5)	100.00	82.54	93.65
(4.5,1.6)	100.00	82.54	93.65
(4.5,1.7)	100.00	80.95	93.98
(4.5,1.8)	100.00	80.95	93.31
(4.5,1.9)	100.00	79.37	93.31
(5.0,1.0)	100.00	87.30	94.31
(5.0,1.1)	100.00	87.30	94.65
(5.0,1.2)	100.00	87.30	94.65
(5.0,1.3)	100.00	87.30	94.31
(5.0,1.4)	100.00	87.30	93.98
(5.0,1.5)	100.00	82.54	93.65
(5.0,1.6)	100.00	82.54	93.65
(5.0,1.7)	100.00	80.95	93.98
(5.0,1.8)	100.00	80.95	93.31
(5.0,1.9)	100.00	79.37	93.31

Because understanding the specificity of the HIV protease is essential for developing inhibitors of the enzyme, and the attempt to define protease inhibitors represents a considerable effort in search for drugs against AIDS, the progresses of the relevant prediction algorithm will improve our ability to reach the goal of finding drugs against AIDS.

Table 6. Comparison to Neural Network Method.

	Method	Rate (%)
Self-consistency	SVM	100
	Neural network	100
Jackknife test	SVM	95
	Neural network	90
Prediction	SVM	87
	Neural network	92

It should be noted that the SVMs method describe here is general and can also be used to predict the substrate specificity of other multisite enzymes, such as GalNAc-transferase (Cai and Chou, in preparation).

References

1. Kohl, N. E.; Emini, E. A.; Schlief, W. A.; Davis, L. J.; Heimbach, J.; Dixon, R. A. F.; Scolnik, E. M.; Sigal, I. S. Proc Natl Acad Sci USA 1988, 85, 4686.
2. Seelmeier, S.; Schmidt, H.; Turk, V.; von der Helm, K. Proc Natl Acad Sci USA 1988, 85, 6612.
3. Hellen, C. U. T.; Krausslich, H. G.; Wimmer, E. Biochemistry 1989, 28, 9881.
4. McQuade, T. J.; Tomasselli, A. G.; Liu, L.; Karacostas, V.; Moss, B.; Sawyer, T. K.; Heinrikson, H. L.; Tarpley, W. J. Science 1990, 247, 454.
5. Graves, B. J.; Hatada, M. H.; Miller, J. K.; Graves, M. C.; Roy, S.; Cook, C. M.; Krohn, A.; Martin, J. A.; Roberts, N. A. In Structure and Function of the Aspartic Protease: Genetics, Structure and Mechanisms; Dunn, B., Ed.; Plenum: New York, 1992; p. 455.
6. Tomasselli, A. G.; Hui, J. O.; Sawyer, T. K.; Thaisrivongs, S.; Hester, J. B.; Heinrikson, R. L. In Structure and Function of Aspartic Proteases; Dunn, B. M., Ed.; Plenum Press: New York, 1991; p. 469.
7. Henderson, L. E.; Benveniste, R. E.; Sowder, R. C.; Copeland, T. D.; Schutz, A. M.; Oroszlan, S. J Virol 1988, 62, 2587.
8. Putney, S. Trends Biotechnol 1992, 17, 191.
9. Chou, J. J. Biopolymers 1993, 33, 1405.
10. Chou, J. J. J Protein Chem 1993, 12, 291.
11. Chou, K. C.; Zhang, C. T. J Protein Chem 1993, 12, 709.
12. Chou, K. C. Anal Biochem 1996, 233, 1.
13. Miller, M.; Schneider, J.; Sathyanarayana, B. K.; Toth, M. V.; Marshall, G. R.; Clawson, L.; Selk, L.; Kent, S. B. H.; Wlodawer, A. Science 1989, 246, 1149.
14. Poorman, R. A.; Tomasselli, A. G.; Heinrikson, R. L.; Kezdy, F. J. J Biol Chem 1991, 266, 14554.
15. Chou, K. C.; Zhang, C. T.; Kezdy, F. J. P. Proteins Struct Funct Genet 1993, 16, 195.
16. Chou, K. C. J Biol Chem 1993, 268, 16938.
17. Zhang, C. T.; Chou, K. C. Protein Eng 1994, 7, 65.
18. Chou, K. C.; Tomasselli, A. G.; Reardon, I. M.; Heinrikson, R. L. Proteins Struct Funct Genet 1996, 24, 51.
19. Vapnik, V. N. The Nature of Statistical Learning Theory; Springer: Berlin, 1995.
20. Vapnik, V. Statistical Learning Theory; Wiley–Interscience, New York, 1998.

21. Burbidge, R.; Trotter, M.; Holden, S.; Buxton, B. Proceedings of the AISB'00 Symposium on Artificial Intelligence in Bioinformatics, 2000, p. 1.
22. Joachims, T. Proceedings of the European Conference on Machine Learning; Springer: Berlin, 1998.
23. Joachims, T. In Making Large-Scale SVM Learning Practical. Advances in Kernel Methods—Support Vector Learning; Schölkopf, B.; Burges, C.; Smola, A., Eds.; MIT Press: Cambridge, MA, 1999.
24. Joachims, T. International Conference on Machine Learning (ICML), 1999.
25. Cortes, C.; Vapnik, V. Mach Learn 1995, 20, 273.
26. Chou, K. C.; Zhang, C. T. Crit Rev Biochem Mol Biol 1995, 30, 275.
27. Zhou, G. P. J Protein Chem 1998, 17, 729.
28. Cai, Y. D. Proteins Struct Funct Genet 2001, 43, 336.
29. Cai, Y. D.; Chou, K. C. Adv Eng Software 1998, 29, 119.