Linear Support Vector Machines for HIV-1 Protease Site Detection

Murat Gök
Science Institute, Sakarya University, Sakarya, Turkey
muratgok@gmail.com

Ahmet Turan Özcerit
Computer Systems Education, Sakarya University, Sakarya, Turkey
aozcerit@sakarya.edu.tr

Abstract: Several studies have been done for the HIV-1 protease specificity problem by applying machine learning computation techniques recently. In this work, a Linear Support Vector Machine (LSVM) technique has been applied to predict the cleavability of proteins by HIV-1 protease. We used Orthonormal Encoding (OE) extraction technique to map octopeptide sequence inputs. According to simulation outcomes, we have achieved better result, which has a rate of %91.8, compared to earlier studies to predict the cleavability of HIV-1 protease.

1. Introduction

HIV-1 protease (Beck et al. 2000) is a small enzyme in the AIDS virus that ensures its replication. The mature and infectious viral particles can only be generated when the polyproteins are cleaved by the HIV protease; otherwise, the viral particles are inactive (Graves et al. 1992). The cleavage sites in the viral polyproteins do not share sequence property. HIV-1 protease inhibitors target the active site in HIV-1 protease for preventing its maturation functioning. Owing to the fact that HIV-1 protease cleavage sites in proteins are templates for inhibitor drugs, deciphering and understanding HIV-1 protease is vital. In this paper, we apply Linear Support Vector Machines (LSVM) for detecting HIV-1 protease cleavage sites in proteins with the view of Orthonormal Encoding technique.

During the last decade or so, for the HIV-1 protease, there have been several works to develop various prediction methods based on machine learning. In (Cai et al. 1998) multilayer perceptron (MLP) which is a non-linear machine learning method was used to solve this problem. Recently, Support Vector Machine (SVM) has been adopted for the prediction of the cleavage sites. In (Cai et al. 2002), the authors applied Vapnik's SVM to the HIV-1 cleavage problem. In (Röngnvaldsson et al. 2003), the authors showed that HIV-1 protease cleavage prediction is a linear problem and LSVM is the best classifier for this problem. In (Nanni et al. 2006), the authors studied an encoding technique that combines the amino acid substitution matrix Blosum50 together with the sequence order of the amino acid composition. In this work, a linear discriminate classifier (LDC) and a radial basis function SVM were combined.

2. Background and Techniques

2.1. Peptide Sequences

A protein sequence is composed of variable combinations of 20 natural amino acids $\sum$(A, C, D, ..., W, Y). A peptide is represented by $P = P_4P_3P_2P_1{\downarrow}P_1{\cdot}P_2{\cdot}P_3{\cdot}P_4$, where $\downarrow$ denotes a scissile bond, P_i is an amino acid belonging to $\sum$ (Röngnvaldsson et al. 2003).

2.2. Orthonormal Encoding (OE) Feature Extraction Technique

Feature extraction is a process that extracts a set of features from the original pattern representation. Orthonormal Encoding is a robust feature extraction technique to map the octopeptide to a sparse orthonormal representation. Each amino acid in the peptide sequence (P_i) is represented by a 20 bits vector with 19 bits set to zero and one bit set to one. We implemented OE in OSU toolbox (http://sourceforge.net/projects/svm/) in MatLab program.

2.3. Data Set

A large data set comprised 1625 substrates has been used. This data set had been experimentally tested for cleavage by the wild type HIV-1 protease (Kontijevski et al. 2007). On this dataset, we conducted 10 tests. Each training and test sets has been randomly re-sampled but maintained the distribution of the patterns in two classes.

2.4. Hyperplane Classifiers: 1-Norm Support Vector Machines (Fung et al. 2005)

Support Vector Machines are a set of related supervised learning methods used for classification and regression. We consider the problem of classifying m points in the n-dimensional input space R^n, represented by the m x n matrix A, according to membership of each point Ai in the class A+ or A− as specified by a given m x m diagonal matrix D with plus ones or minus ones along its diagonal. For this problem, depicted in Figure 1, the linear programming support vector machine with a linear kernel (this is a variant of the standard SVM) is given by the following linear program with parameter $\nu > 0$:

$$\min_{(\omega,\gamma,y)\in R^{n+1+m}} \nu e'y + \|\omega\|_1$$
$$\text{s.t. } D(A\omega - e\gamma) + y \geq e \qquad (1)$$
$$y \geq 0,$$

Where $\|.\|_1$ denotes the 1-norm as defined in the Introduction. That problem is indeed a linear program, and it can be easily seen from the equivalent formulation below:

$$\min_{(\omega,\gamma,y)\in R^{n+1+m}} \nu e'y + e't$$
$$\text{s.t. } D(A\omega - e\gamma) + y \geq e \qquad (2)$$
$$t \geq \omega \geq -t,$$
$$y \geq 0.$$

For economy of notation we shall use the first formulation (1) with the understanding that computational implementation is via (2). If the classes are linearly inseparable, which is often the case in real-world datasets, then two planes bound the two classes with a "soft margin" (i.e. bound approximately with some error) determined by the nonnegative error variable y, that is:

$$A_i\omega + y_i \geq \gamma + 1, \quad \text{for } D_{ii} = 1, \qquad (3)$$
$$A_i\omega + y_i \geq \gamma - 1, \quad \text{for } D_{ii} = -1.$$

The 1-norm of the error variable y is minimized parametrically with weight ν in (1), resulting in an approximate separating plane. This plane classifies data as follows:

$$sign(x'\omega - \gamma) \begin{cases} = 1, then\ x \in A_+, \\ = -1, then\ x \in A_-, \end{cases} \qquad (4)$$

Where sign $(\cdot)$ is the sign function defined in the Introduction. Empirical evidence indicates that the 1-norm formulation has the advantage of generating very sparse solutions. This results in the normal ω to the separating plane $x'\omega = \gamma$ having many zero components, which implies that many input space features do not play a role in determining the linear classifier.

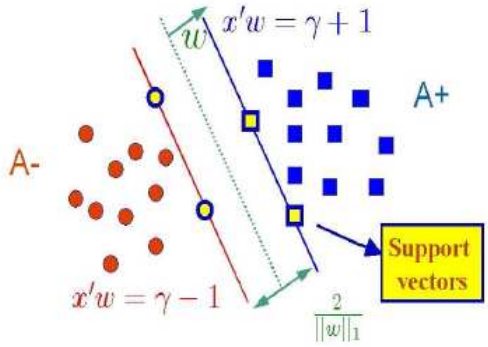

Figure 1. The LP-SVM classifier in the w-space of Rn. The plane of equation (3) approximately separating points in A+ from points in A−.

This makes this approach suitable for feature selection in classification problems. Since our rule extraction algorithm depends directly on the features used by the hyperplane classifier, sparser normal vectors ω will lead to rules depending on a fewer number of features.

3. Results and Discussion

Owing to the fact that OE makes the domain space sparser (i.e. from 20^8 to 2^{160}), the ratio of number of features and the sample size increases. In Figure 2, a graphical description of the system experimented is given. On step 1, training data encodes as Orthonormal encoding. Referred training data is labeled with respect to (Kontijevski 2007). On step 2, By Using orthonormal encoded inputs, LSVM Classifier is constructed. On step 3, testing data is encoded as orthonormal encoding. Finally, a confusion matrix which is a table with two rows two columns that reports the number of True Negatives, False Positives, False Negatives, and True Positives produced from LSVM classifier.

We conducted 10 tests and get 0.9277 class rate values as stand-of-the art. We have used totally 540 samples as seen Table 1.

Data	Training	Test
Cleaved data	184	190
Uncleaved data	356	350
Total	540	540

Table 1. Sample data experimented

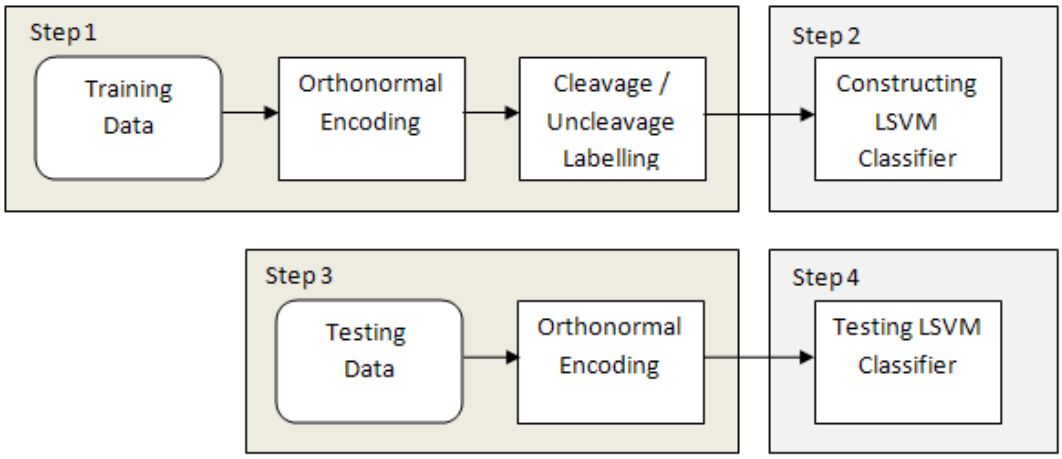

Figure 2. Graphical description of the system experimented

4. Conclusion

In this paper, the problem is to predict whether a given peptide, HIV-1 protease cleavage site, is cleaved or not. We showed by an empirical approach, LSVM with OE inputs can partition the space successfully. Experimental results show that our result outperforms the results as pointed in (Nanni et al. 2006) has a value of 91.8% and in (Kim et al. 2008) has a value of 89.92%.

References

Beck, Z.Q., Hervio, L., Dawson, P.E., Elder, J.E., Madison, E.L. (2000). *Identification of efficiently cleaved substrates for HIV-1 protease using a phage display library and use in inhibitor development*. Virology.

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. (1992). *In Structure and Function of the Aspartic Protease: Genetics, Structure and Mechanisms*, Dunn, B., Ed. Plenum: New York; p. 455.

Cai Y.D., Chou K.C. (1998). *Artificial neural network model for predicting HIV protease cleavage sites in protein.* Adv Eng Software 29:119-128

Cai Y.D., Liu X.J., Xu X.B., Chou K.C. (2002). *Support vector machines for predicting HIV protease cleavage sites in protein.* JComputer Chemistry 23: 267–274.

Röngnvaldsson, T., You, L. (2003). *Why neural networks should not be used for HIV-1 protease cleavage site prediction.* Bioinformatics, 1702–1709.

Nanni L., Lumini A. (2006). *A reliable method for HIV-1 protease cleavage site prediction methods.* Neuro Computing 69: 838 - 841.

Kontijevski A., Wikberg J.E.S., Komorowski J. (2007). *Computational proteomics analysis of HIV-1 protease interactome*. Proteins:Structure, Function and Bioinformatics 68: 305–312.

Fung G., Sandilya S.,Rao R.B. (2005). *Rule Extraction from Linear Support Vector Machines.* KDD'05, August 21–24, Chicago, Illinois, USA.

Kim H., Zhang Y., Heo Y., Oh H., Chen S. (2008). *Specificity rule discovery in HIV-1 protease cleavage site analysis.* Compuatational Biology and Chemistry 32: 72–79.