International Journal of
Intelligent Systems and Applications in Engineering

ISSN:2147-6799 www.ijisae.org Original Research Paper

# Comparison among Feature Encoding Techniques for HIV-1 Protease Cleavage Specificity

**Uğur Turhal[1], Murat Gök*[2], Aykut Durgut[3]**

*Accepted 15th August 2014* **DOI: 10.18201/ijisae.21005**

***Abstract:*** HIV-1 protease which is responsible for the generation of infectious viral particles by cleaving the virus polypeptides, play an indispensable role in the life cycle of HIV-1. Knowledge of the substrate specificity of HIV-1 protease will pave the way of development of efficacious HIV-1 protease inhibitors. In the prediction of HIV-1 protease cleavage site techniques, many efforts have been devoted. Last decade, several works have approached the prediction of HIV-1 protease cleavage site problem by applying a number of methods from the field of machine learning. However, it is still difficult for researchers to choose the best method due to the lack of an effective and up-to-date comparison. Here, we have made an extensive study on feature encoding techniques for the problem of HIV-1 protease specificity on diverse machine learning algorithms. Also, for the first time, we applied OEDICHO technique, which is a combination of orthonormal encoding and the binary representation of selected 10 best physicochemical properties of amino acids derived from Amino Acid index database, to predict HIV-1 protease cleavage sites.

***Keywords:*** *HIV-1 protease specificity, Feature extraction, Peptide classification, Machine learning algorithms, Amino acids*

## 1. Introduction

Owing to the fact that HIV-1 protease cleaves virus polypeptides at defined susceptible sites, it is responsible for processing polyproteins that contain the structural proteins (Gag polyprotein) and the enzymes (Gag/Pol polyprotein) required for virus structure and replication (Zachary Q. Beck et al, 2000). The investigation of substrate specificity of HIV-1 protease is the ultimate goal as well as to identify optimal sequences to act as a framework for development of highly efficient inhibitors which target the active site of the protease. Therefore, the knowledge of the polyprotein cleavage sites by HIV-1 protease is vital. The cleavability prediction task, given a sequence of eight amino acids (an octamer), aims at knowing which peptide sequences are cleaved or non-cleaved by the protease. So far, no perfect task is yet known that determines where a peptide will be cleaved or non-cleaved by the protease. A standard feed-forward multilayer perceptron (MLP) was used (You-Dong Cai and Kuo-Chen Chou, 1998, Thomas B. Thompson et al., 1995). Support vector machines (SVM) was adopted several times to predict the cleavability (Yu-Dong Cai et al., 2002, Thorsteinn Rognvaldsson and Liwen You, 2004, Loris Nanni, 2006). In (Thorsteinn Rognvaldsson and Liwen You, 2004), authors showed that the linear SVM works well for the problem. But these all works were conducted on an out-of-date dataset including 362 peptide substrates. More recently, a larger dataset (PR-1625), which was composed of 1625 peptide sequences, was provided by Kontijevskis et al (Aleksejs Kontijevskis et al., 2007). They proposed a statistically valid rule-based model for the HIV-1 protease specificity. In (Bing Niu et al., 2009) k-Nearest Neighbors (kNN) algorithm was applied with a feature representation based on the amino acid properties constructed by Amino Acid index database (AAindex) (Shuichi Kawashima and Minoru Kanehisa, 2000).

For the protease specificity, several studies have been conducted applying orthonormal encoding (OE) method along with different machine learning techniques (You-Dong Cai and Kuo-Chen Chou, 1998, Loris Nanni, 2006). Ruan et al. (Jishou Ruan et al., 2005) proposed a new encoding, termed composition moment vector (CMV), and based residue's order in a protein sequence. Another approach is OETMAP encoding (Murat Gök and Ahmet T. Özcerit, 2012) which is in conjunction of OE and Taylor's Venn-diagram (TVD). OEDICHO feature encoding (Murat Gök and Ahmet T. Özcerit, 2012), which was created for the T-cell epitopes recognition, has been tested for the first time for HIV-1 protease specificity problem in this study.

In this paper, five encoding techniques with five learning algorithms as a standalone approach have been performed to predict HIV-1 protease cleavage sites on PR-1625 HIV-1 dataset. The computational results demonstrate higher performance of OEDICHO in comparison with the feature encoding methods on a standalone classifier approaches.

## 2. Methods

### 2.1. Feature Encoding

Feature encoding, which is a commonly used technique applied before classification, defines a mapping from the original representation space into a new space where the classes are more easily separable. The goal of feature encoding is to distill the pattern data into a more concentrated and manageable form. This will reduce the classifier complexity, increasing in most cases classifier accuracy (Alberto J. Perez-Jimenez and Juan C.

[1,3] *University of Balıkesir – 10100, Turkey*
[2] *University of Yalova – 77100, Turkey*
** Corresponding Author: Email: murat.gok@yalova.edu.tr*
*Note :This paper has been presented at the International Conference on Advanced Technology&Sciences (ICAT'14) held in Antalya (Turkey), August 12-15, 2014.*

Perez-Cortes, 2006). The evaluated feature encoding techniques are given in brief terms as follows:

1) OE technique is a common encoding method. According to OE, each amino acid symbol Pi in a peptide is replaced by an orthonormal vector, $d_i = (\delta_{i1}, \delta_{i2}, ..., \delta_{i20})$ where $\delta_{ij}$ is the Kronecker delta symbol. Then, each $P_i$ is represented by a 20-bit vector, 19 bits are set to zero and 1 bit is set to one based on alphabetic order of amino acids. Each $d_i$ vector is orthogonal to the rest of $d_i$ vectors and $P_i$ can be any one of the twenty amino acids available in human body (Thorsteinn Rognvaldsson and Liwen You, 2004). The main drawback of OE technique is that OE binary feature vectors result in information loss.

2) CMV encoding includes information about both composition and position of amino acids in the sequence. CMV provides functional relation with the structure content, i.e. there must not be two or more primary amino acid sequences that would have different structure content but the same composition moment vector. The feature vector is obtained concatenating the first, the second and the third moment vectors (Jishou Ruan et al., 2005).

3) OETMAP encoding consists of a conjunction of OE and Taylor's Venn-diagram methods which are complementary to each other. OETMAP shows the relationship of the 20 naturally occurring amino acids to a selection of physicochemical properties which are important in the determination of protein tertiary structure (Murat Gök and Ahmet T. Özcerit, 2012)

4) OEDICHO encoding combines the OE and amino acid physicochemical properties knowledge to give complementary recognition information to OE. It therefore increases the effectiveness of OE. 10 best physicochemical properties were selected with pc-based encoding and were converted their index's values into binary feature vectors. Then OE and this complementary binary feature vector were concatenated for each residue in the peptide sequence (Murat Gök and Ahmet T. Özcerit, 2012).

### 2.2. Classification

A protein sequence is composed of a series of 20 amino acids represented by characters as A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y and V. A peptide is denoted by P = P4P3P2P1↓P1'P2'P3'P4' where Pi is an amino acid. The scissile bond is located between positions P1 and P1' (Israel Schechter and Arieh Berger, 1967). The HIV-1 protease cleavage specificity can be considered as a binary classification problem where an octomer peptide is required to be assigned to cleavable or non-cleavable class. We used five types of learning algorithms such as J48, kNN, K-star, linear SVM and MLP for classification under WEKA data mining software. The evaluated algorithms are given in brief terms as follows:

1) J48 classifier is a simple C4.5 algorithm. C4.5 generates classifiers expressed as decision trees. In decision tree learning, the learned function is represented by a decision tree. Each node in the tree represents a test of some attribute of the training example, and each branch corresponds to one of the possible values for its source node (attribute) (Pang-Ning Tan et al., 2005). J48 creates a binary tree.

The confidence factor parameter which is used for pruning has been set to 0.25.Minimum number of instances per leaf has been chosen as 2.

2) kNN classification, finds a group of k objects in the training set that are closest to the test object, and bases the assignment of a label on the predominance of a particular class in this neighborhood. There are three key elements of this approach: a set of labeled objects, a distance or similarity metric to compute distance between objects, and the value of k, the number of nearest neighbors. To classify an unlabeled object, the distance of this object to the labeled objects is computed, its k-nearest neighbors are identified, and the class labels of these nearest neighbors are then used to determine the class label of the object (Xindong Wu et al., 2008).

3) KStar is an instance-based classifier. The class of a test instance is based upon the class of those instances in the training set similar to it, as determined by some similarity measurement. The underlying assumption of this type of classifiers is that similar cases belong to the same class (Du Zhang and Jeffrey J. P. Tsai, 2007).

The global blend parameter for KStar has been chosen as 20.

4) SVM is an effective discriminative classification method of statistic learning theory and in recent times, it is successively applied by a number of other researchers. SVM aims to find the maximum margin hyperplane to separate two classes of patterns. In linear SVM classifier, classes of two patterns are linearly separated by the use of a maximum margin hyperplane, which is uniquely defined by the support vectors, gives the best separation between the classes (Christopher J.C. Burges, 1998). The parameters of the stand-alone SVM (kernel ∈{'linear'}, cost of the constrain violation (C) = 10) have been applied for the dataset.

5) Multiple layer perceptron (MLP) are the most common type of feed-forward networks and the neurons are organized into layers that have unidirectional connections between them. Based on this rationale, the learning process can be viewed as the problem of updating connection weights from available training patterns so that the network can efficiently perform a specific task. Performance is improved over time by iteratively updating the weights in the network. This learning process can be efficiently performed with the Back-propagation algorithm, which is based on gradient descend. Then, the actual output of the network is generated and the (possible) error produced by the difference between the actual output and the desired output is used to modify the connections weights in order to gradually reduce the overall error (David E. Rumelhart et al., 1986).

The parameters of MLP were learning rate = 0.3, momentum = 0.2.

## 3. Results

### 3.1. Experimental Setup

We conducted our tests on PR-1625 (Aleksejs Kontijevskis et al., 2007) HIV-1 protease dataset, composed by sequences of eight amino acids that can be cleaved or not by the HIV-1 protease. PR-1625 contains 1625 octamer peptides, of which 374 are cleaved and 1251 are non-cleaved.

Given a set of peptide sequences with known labels of cleavable and non-cleavable, we built OEDICHO encoding feature vectors. OEDICHO was implemented as in (Murat Gök and Ahmet T. Özcerit, 2012). To create binary feature vector for HIV-1 protease site prediction, firstly, 10 best physicochemical properties were selected according to physicochemical (pc) encoding technique and were converted their index's values into binary feature vectors. Linear support vector machines were used as the classifier in case of being implementing pc-based encoding. Pc-based encoding was repeated for each 544 physicochemical properties indices in AAindex (Murat Gök and Ahmet T. Özcerit, 2012). According

**Table 1.** 10-best physicochemical properties selected for PR-1625

| Order | Physicochemical Property (PR-1625) |
|---|---|
| 1 | Energy transfer from out to in (95%buried) |
| 2 | Side chain interaction parameter |
| 3 | Average relative fractional occurrence in AL(i) |
| 4 | Average gain in surrounding hydrophobicity |
| 5 | Hydrophobicity-related index |
| 6 | Membrane preference for cytochrome b: MPH89 |
| 7 | ALTLS index |
| 8 | Hydrophobicity scale from native protein structures |
| 9 | Ratio of buried and accessible molar fractions |
| 10 | Normalized frequency of turn in alpha+beta class |

In the second step, for each best property's mean was calculated. If the index value is smaller than or equals the mean value, it is accepted as 0. If the index value is bigger than the mean value, it is accepted as 1. In this way, binary encoding table was obtained. Consequently, different pc-encoding tables obtained for each dataset.

The built feature vectors, OE (20-bit) and binary encoding for the best 10 physicochemical property of each residue (10-bit) is concatenated, respectively. Hence, the feature vector has a dimension of 240-bit (30-bit x 8 residue) for each octomer peptide.

The performances of the classifiers were evaluated by means of accuracy (acc), kappa error (ke), F-score and the Matthews correlation coefficient (MCC) value metrics. True positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) values are obtained via confusion matrix (Tom Fawcett, 2004, Jin Huang and Charles X. Ling, 2005). Acc is a widely used measure to determine class discrimination ability, and it is calculated as:

$$acc = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

In statistical analysis, quantifying the performance of classifier to accuracy performance obtained, 10-best physicochemical properties, shown in Table 1, were selected for PR-1625. algorithms in terms of just using the percentage of misses as the single meter for accuracy can give misleading results. Therefore, the cost of error must also be taken into account. Kappa error is a good metric to inspect classifications that may be due to chance. Kappa error takes values between (−1, 1). As the Kappa value approaches to 1, then the performance of the related classifier is assumed to be more realistic rather than by chance. Kappa error is calculated as (Arie Ben-David, 2008):

$$ke = \frac{p_0 - p_c}{1 - p_c} \quad (2)$$

In Eq. 2, is the total agreement probability and $p_c$ is the agreement probability due to chance.

MCC, which is used as a measure of the quality of binary (two-class) classifications, takes into account true and false positives and negatives.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (3)$$

F-score is a measure of a test's accuracy determining accuracy accounting for both precision and for recall from confusion matrix. F-score accounted as shown in Eq. 4.

$$F = 2 \times \frac{precision \times recall}{precision + recall} \quad (4)$$

### 3.2. Performance of Feature Encoding Techniques

10-fold cross validation (10-fold CV) testing scheme is applied to evaluate the performance of the methods in terms of accuracy, kappa, F-score and averaged over 10 experiments on PR-1625. In a cross-validation run, the 10 folds are randomly created. In 10-fold CV, the encoding scheme methods are trained using 90 % of the data and the remaining 10 % of the data are used for testing of the methods (Richard O. Duda et al., 2000). This process is repeated 10 times so that each peptide in datasets is used once. The 10 folds used in the training are different from the 10 folds used in the testing. Having completed the procedures above, the average accuracy, kappa statistic, F-score and MCC values of the each method over these 10 turns are obtained, as shown Table 2 and Table 3.

**Table 3.** F-Score and MCC Performances of All Methods on Pr-1625 HIV-1 Dataset

| | Linear SVM | | MLP | | J48 | | IBK | | K-Star | |
|---|---|---|---|---|---|---|---|---|---|---|
| | F-Score | MCC | F-Score | MCC | F-Score | MCC | F-Score | MCC | F-Score | MCC |
| **OE** | 0,94 | 0,84 | 0,94 | 0,84 | 0,91 | 0,76 | 0,90 | 0,75 | 0,92 | 0,79 |
| **CMV** | 0,79 | 0,47 | 0,89 | 0,69 | 0,90 | 0,73 | 0,89 | 0,69 | 0,94 | 0,83 |
| **OE+CMV** | 0,90 | 0,72 | 0,87 | 0,64 | 0,92 | 0,76 | 0,88 | 0,66 | 0,89 | 0,71 |
| **OETMAP** | 0,95 | 0,85 | 0,95 | 0,86 | 0,91 | 0,75 | 0,90 | 0,74 | 0,91 | 0,76 |
| **OEDICHO** | 0,95 | 0,85 | 0,96 | 0,88 | 0,92 | 0,77 | 0,90 | 0,73 | 0,90 | 0,74 |

The results points out that OEDICHO has obtained the best result for both F-score and MCC values with the value of 0.96 and 0.88, respectively.

**Table 2.** Accuracy Performance of Amino Acid Encoding Schemes on Pr-1625 HIV-1 Dataset

| | Linear SVM | | MLP | | J48 | | IBK | | K-Star | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Acc (%) | Kappa | Acc (%) | Kappa | Acc (%) | Kappa | Acc (%) | Kappa | Acc (%) | Kappa |
| **OE** | 94.34 | 0.84 | 94.22 | 0.84 | 91.45 | 0.76 | 90.03 | 0.74 | 92.06 | 0.77 |
| **CMV** | 83.14 | 0.37 | 89.17 | 0.69 | 90.52 | 0.73 | 89.29 | 0.69 | 94.22 | 0.83 |
| **OE+CMV** | 90.52 | 0.70 | 88.12 | 0.63 | 91.57 | 0.76 | 87.26 | 0.66 | 88.62 | 0.70 |
| **OETMAP** | 94.65 | 0.85 | 95.20 | 0.87 | 91.08 | 0.75 | 89.60 | 0.73 | 90.52 | 0.76 |
| **OEDICHO** | 94.52 | 0.85 | 95.63 | 0.88 | 92.00 | 0.77 | 89.17 | 0.72 | 89.85 | 0.73 |

The results report that OEDICHO encoding outperforms the competing encoding techniques with the value of 95.63 %. Also highest Kappa value of OEDICHO with MLP (0.88) confirms its robustness of measurement. Note that IBK algorithm obtained the worst performance among learning algorithms with nearly all feature encoding techniques.

OEDICHO method yields high accurate empirical results compared to all feature encoding techniques for HIV-1 protease specificity. We notice that OEDICHO combines the both effectiveness of OE and selected 10-best physicochemical properties of AAindex. However, OETMAP uses TVD that is not up-to-date in comparison with OEDICHO's 10-best physicochemical properties. A rule based method, the orthogonal search-based rule extraction (OSRE), is presented by Rognvaldsson et al. (Thorsteinn Rögnvaldsson et al., 2009). OSRE obtained accuracy of 93% on PR-1625. OEDICHO shows better performance with the accuracy of 95.63 %.

## 4. Conclusions

The problem addressed in this paper is to recognize, given a sequence of amino acids, HIV-1 protease cleavage site. We performed an experimental comparison of five encoding technique with five learning classifiers such as linear SVM, MLP, J48, IBK and K-Star using up-to-date PR-1625 HIV-1 dataset. OEDICHO technique, which joins OE technique and complementary binary feature vector which comprises selected 10 best physicochemical properties' index values for each residue in a peptide sequence, shows the best performance with MLP algorithm. Beside HIV-1 protease site prediction, OEDICHO encouraged to be used for other peptide classification problems. Due to the fact that independent and accurate classifiers can make errors on different regions of the feature space, they can be ensemble to achieve better scores. Based on this rationale, future works might involve the ensemble of classifiers with diverse encoding techniques especially with OEDICHO encoding.

## Acknowledgment

This work was supported by Yalova University, YL Project (Grant 2014/YL/037).

## References

[1] Zachary Q. Beck, Laurence Hervio, Philip E. Dawson, John H. Elder and Edwin L. Madison (2000). Identification of Efficiently Cleaved Substrates for HIV-1 Protease Using a Phage Display Library and Use in Inhibitor Development. Virology. Vol. 274. Pages. 391-401.

[2] You-Dong Cai and Kuo-Chen Chou (1998). Artificial Neural Network Model for Predicting HIV Protease Cleavage Sites in Protein. Advances in Engineering Software. Vol. 29. Pages. 119-128.

[3] Thomas B. Thompson, Kuo-Chen Chou and Chong Zheng (1995). Neural Network Prediction of the HIV-1 Protease Cleavage Sites. Journal of Theoretical Biology. Vol. 177. Pages. 369–379.

[4] Yu-Dong Cai, Xiao-Jun Liu, Xue-Biao Xu and Kuo-Chen Chou (2002). Support Vector Machines for Predicting HIV Protease Cleavage Sites in Protein. Journal of Computational Chemistry. Vol. 23. Pages. 267-274.

[5] Thorsteinn Rognvaldsson and Liwen You (2004). Why Neural Networks Should Not Be Used For HIV-1 Protease Cleavage Site Prediction. Bioinformatics. Vol. 20. Pages. 1702-1709.

[6] Loris Nanni (2006). Comparison among Feature Extraction Methods for HIV-1 Protease Cleavage Site Prediction. Pattern Recognition. Vol. 39. Pages. 711-713.

[7] Aleksejs Kontijevskis, Jarl E. S. Wikberg and Jan Komorowski (2007). Computational Proteomics Analysis of HIV-1 Protease Interactome. Proteins: Structure, Function and Bioinformatics. Vol. 68. Pages. 305-312.

[8] Bing Niu, Lin Lu, Liang Liu, Tian H. Gu, Kai-Yan Feng, Wen-Cong Lu and Yu-Dong Cai (2009). HIV-1 Protease Cleavage Site Prediction Based on Amino Acid Property. Journal of Computational Chemistry. Vol. 30. Pages. 33-39.

[9] Shuichi Kawashima and Minoru Kanehisa (2000). AAindex: Amino Acid Index Database. Nucleic Acids Research. Vol. 28. Page. 374. Available online: http://www.genome.jp/dbget/aaindex.html

[10] Jishou Ruan, Kui Wang, Jie Yang, Lukasz A. Kurgan and Krzysztof Cios (2005). Highly Accurate and Consistent Method for Prediction of Elix and Strand Content from Primary Protein Sequences. Artificial Intelligence in Medicine. Vol. 35. Pages. 19-35.

[11] Murat Gök and Ahmet T. Özcerit (2012). OETMAP: A New Feature Encoding Scheme for MHC Class I Binding Prediction. Molecular and Cellular Biochemistry. Vol. 359. Pages. 67-72.

[12] Murat Gök and Ahmet T. Özcerit (2012). Prediction of MHC Class I Binding Peptides with a New Feature Encoding Technique. Cellular Immunology. Vol. 275. Pages. 1-4.

[13] Alberto J. Perez-Jimenez and Juan C. Perez-Cortes (2006). Genetic Algorithms for Linear Feature Extraction. Pattern Recognition Letters. Vol. 27. Pages. 1508-1514.

[14] Israel Schechter and Arieh Berger (1967). On the Size of the Active Site in Proteases. I. Papain. Biochemical and Biophysical Research Communications. Vol. 27. Pages. 157-162.

[15] Pang-Ning Tan, Michael Steinbach and Vipin Kumar (2005). Introduction to Data Mining. Addison-Wesley Longman Publishing Co., Inc. Boston, ISBN: 0321321367.

[16] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep

Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach, David J. Hand and Dan Steinberg (2008). Top 10 Algorithms in Data Mining. Knowledge and Information Systems. Vol. 14. Pages. 1-37.

[17] Du Zhang and Jeffrey J. P. Tsai (2007). Advances in Machine Learning Applications in Software Engineering. Idea Group Publishing. Page. 253.

[18] Christopher J.C. Burges (1998). A Tutorial on Support Vector Machines for Pattern Recognition. Data Mining and Knowledge Discovery. Vol. 2. Pages. 121-167.

[19] David E. Rumelhart, Geoffrey E. Hinton and Ronald J. Williams (1986). Learning Internal Representations by Error Propagation. Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Vol. 1. Pages. 318-362.

[20] Tom Fawcett (2004). ROC Graphs: Notes and Practical Considerations for Researchers. Technical Report, HP Laboratories. MS 1143. 1501 Page Mill Road. Palo Alto CA 94304. USA, 2004.

[21] Jin Huang and Charles X. Ling (2005). Using AUC and Accuracy in Evaluating Learning Algorithms. IEEE Transactions on Knowledge and Data Engineering. Vol. 17. Pages. 299-310.

[22] Arie Ben-David (2008). Comparison of Classification Accuracy Using Cohen's Weighted Kappa. Expert Systems with Applications. Vol. 34. Pages. 825-832.

[23] Richard O. Duda, Peter E. Hart and David G. Stork (2000). Wiley. New York. ISBN: 978-0-471-05669-0.

[24] Thorsteinn Rögnvaldsson, Terence A. Etchells, Liwen You, Daniel Garwicz, Ian Jarman and Paulo J. G. Lisboa (2009). How to Find Simple and Accurate Rules for Viral Protease Cleavage Specificities. BMC Bioinformatics. Vol. 10. Page. 149.