CHAPTER IV

HIV-1 CRF01_AE Coreceptor Usage Prediction using Kernel Methods based Logistic Model Tree

4.1 Introduction

The determination of HIV-1 coreceptor usage can be divided into phenotypic and genotypic assays. Phenotypic assays are based on a pseudotyped virus or a recombinant virus which is the gold standard method for determination of the HIV-1 coreceptor usage. This approach is considered to be the most reliable and accurate [73], but it is a very complex, time-consuming, laborious and expensive process, and it also requires special facilities. Thereafter, the third variable loop (V3) of the gp120 envelope as the major domain associated with coreceptor usage is determined [74]. For the Genotypic assay, several genotypic or computational methods have been developed to predict coreceptor usage from the amino acid sequence of V3.

There are various computational methods based on the V3 amino acid sequences. The basic method is the 11/25 rule that is based on a basic amino acid (R or K) at either the 11 or 25 position of the V3 region which is associated with the CXCR4 coreceptor usage [78], whereas acidic or neutral amino acids at these positions are associated with CCR5 coreceptor usage. Using the net charge rule, it was proposed that a net charge of > +5 for the V3 loop (calculated by subtracting the number of negatively charged amino acids [D and E] from the number of positively charged ones [K and R]) is associated with CXCR4 coreceptor usage, whereas a net charge of <+5 is associated with CCR5

Table 4.1 Some existing methods for predicting HIV-1 coreceptor usage

Reference	Classifier	Dataset		features	#feature	feature selection
		training	testing			
Resch et al, 2001 [76]	NN	181	-	V3 amino acid sequence	35	No
Pillai et al, 2003 [75]	SVM	271	-	V3 amino acid sequence	35	No
Jensen et al, 2003 [77]	Threshold	213	175	PSSM	400	No
Jensen et al, 2006 [82]	Threshold	279	-	PSSM	400	No
Sander et al, 2007 [83]	SVM	432	-	structure descriptions: (5 functional atom types)	5	No
Lamers et al, 2008 [84]	NN	127	30	exchange group, charge polarity, hydrophobicity, mass,structural,2D propensity	250	evolved NN
Boisvert et al, 2008 [85]	SVM	1425	1425	similarity of sequence (string kernel)	1425	No
Our work [2]	**SVM**	**273**	**-**	**V3 amino acid sequence**	**9**	**LMT**

coreceptor usage [78,79]. For the Combined rule, a sample was defined as an X4-virus if any of the 11/25 or net charge rule classified it as an X4, otherwise it is an R5-virus [80]. Several bioinformatics tools have been developed as genotypic predictors. The more details about computational methods and bioinformatics tools were summarized in Table 4.1.

Most existing genotypic predictor tools including G2P have been designed based on the genetic characterization and genotype-phenotype correlations of HIV-1 subtype B. Therefore, using such genotypic tools is mainly reliable for B viruses. However, for non-B subtypes, these tools might fail to detect X4 variants [81]. Since the CRF01_AE of HIV-1 is the predominant circulating subtype found in Thailand and Southeast Asia, there is an urgent need to know the performance of genotypic predictor tools for a prediction of coreceptor usage in this subtype.

Our objective in this study is to develop computational approaches to obtain the concordant model in the CRF01_AE subtypes and also to improve the model based on our combination method. In this study, we introduce an approach using the support vector machine method (SVM) in conjunction with the logistic model tree (LMT), which is used as the feature selector [2]. Since, SVM is a very popular and method deal to with classification problem. And LMT gives more accurate results and also outperforms well-known enhanced decision tree learners.

4.2 Dataset

We obtained a total of 273 amino acid sequences for the V3 region of HIV-1 gp120 from the HIV database (http://www.hiv.lanl.gov/ components/sequence/HIV/ search/search.html) which originated from 273 isolations from individual patients. The setting for the V3 region of the CRF01_AE clade and CCR5 or CXCR4 coreceptors was set as the search criteria. The result consisted of 177 isolations showing the R5 phenotype, and 96 showing the X4 phenotype. These sequences were then used to identify coreceptor usage and evaluate the concordance among the available bioinformatics tools.

4.3 Computational methods for HIV-1 CRF01_AE Coreceptor Usage Prediction

To predict HIV-1 CRF01_AE coreceptor usage, we proposed the combination of the SVM classifier and LMT method for improving the predictive result. More details of the SVM classifier and LMT method can be found in Section 2.4 and 2.6, respectively.

4.3.1 Implementation using kernel methods based LMT

For the features which are selected by using the LMT, more details can be referred in Chapter 2, which show the most relevant attributes in our data. In the case of the LMT method, logistic regression is performed using a simple regression (to select informative features). The system flowchart of the SVM classifier based LMT method is shown in Figure. 4.1. To build feature sets with a high prediction potential we applied an algorithm that adapted the idea of LMT for the classification problems using logistic regression instead of linear regression at the leaves. The most relevant attributes consist of 9 positions on V3 amino acid sequences, i.e. P_5 , P_7 , P_{11} , P_{13} , P_{14} , P_{18} , P_{19} , P_{27} , P_{32}.

In this work, we focused on the prediction of the HIV-1 coreceptor which was related to the problem of binary-class classification. The SVM classifier is a well-known method used in such problems. We used a binary SVM classifier to classify the coreceptor into two classes, i.e. CCR5 and CXCR4. In addition, we exploited the idea of feature selection to construct an efficient SVM classifier by using only informative features. In the classification problem with a training data set *D*, we predicted the class of unknown data x_{n+1} by using a linear decision function determined by the kernel function of the inner product between feature vectors. The decision function was then expressed as shown in Eq. 2.9. Recall in Chapter 2, the $K(\cdot,\cdot)$ is the kernel function. In our experiment we used the prime kernel functions, i.e. linear kernel ($K_{linear}(x, x')$), polynomial kernel ($K_{poly}(x, x')$) with degrees d = 2, 3, and 4, and also a radial basis function ($K_{RBF}(x, x')$) kernel. The typical kernel functions are the following:

$$K_{linear}(x, x') = \Phi(x)^{\mathrm{T}}\Phi(x^{'}) \tag{4.1}$$

$$K_{poly}(x, x') = (1 + \gamma \cdot \Phi(x)^{\mathrm{T}}\Phi(x^{'}))^{d} \tag{4.2}$$

$$K_{RBF}(x, x') = \exp(\gamma \cdot \left\|\Phi(x) - \Phi(x^{'})\right\|) \tag{4.3}$$

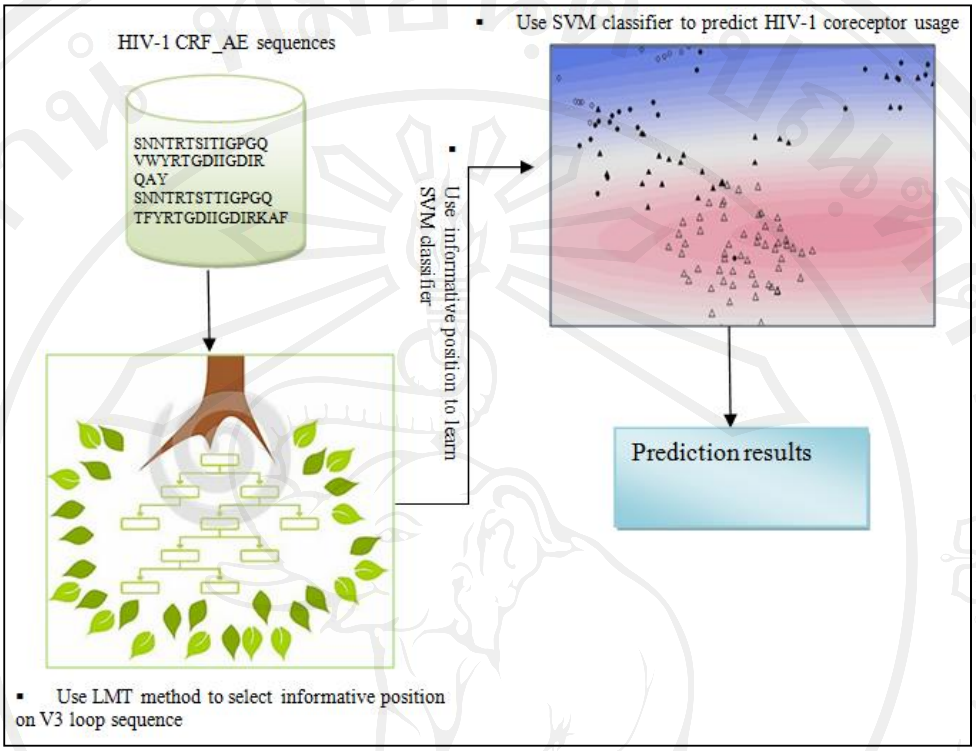
HIV-1 CRF_AE sequences
SNNTRTSITIGPGQ
VWYRTGDIIGDIR
QAY
SNNTRTSTTIGPGQ
TFYRTGDIIGDIRKAF
▪ Use SVM classifier to predict HIV-1 coreceptor usage
▪ Use informative position to learn SVM classifier
Prediction results
▪ Use LMT method to select informative position on V3 loop sequence

Figure 4.1 - The system flowchart of the SVM classifier based LMT method to predict HLA-1 CRF_AE coreceptor usage.

4.4. Description of the Experiments

In this work, the classes of the HIV-1 coreceptor of subtype CRF01_AE were classified by the SVM classifier based LMT method. Given a sequence of observations$x = (x_1, x_2, \ldots, x_n)$, the class of the HIV-1 coreceptor $y = (y_1, y_2, \ldots, y_n)$ is obtained using the SVMs, where $y_i \in \{$ CCR5, CXCR4$\}$.

The selected data set was randomly split into 10 subsets of roughly the same size. During each 10-fold cross-validation experiment, 9 subsets were used for training and the remaining subset was used for validation. Finally, the results were then averaged across the 10 cross-validation experiments. The predictive performances were evaluated for accuracy (ACC), specificity (Spec), and sensitivity (Sens). The Kernel Support Vector Machine (KSVMs) was selected to classify the HIV-1 coreceptor. The package used in this thesis was the KSVMs in RGui which is part of kernlab [67]. To perform the KSVMs we used a linear kernel, polynomial kernel with $d = 2,3,4$ and the radial basis function kernel were evaluated. Moreover, we implemented the LMT to select the most relevant features in our dataset. The LMT was used to improve the SVM classifier by identifying and removing the irrelevant and redundant features. The package of the LMT used here is called theRGui which is Weka_classifier_trees [86], and can be downloaded freely from http://cran.project.org/ web/packages/RWeka /index.html.

4.5. Results and discussion

4.5.1 Results and discussion of kernel methods based LMT

A ten-fold cross validation was performed to classify 273 amino acid sequences into CCR5 or CXCR4 using a binary SVM. In Table 4.2, we compare the predictive performances of this SVM using different kernel functions and features. Each of the five major rows is a different kernel, i.e. linear, linear kernel, polynomial kernels with d = 2, 3, and 4 and the radial basis function. To classify the HIV-1 coreceptor usage, each kernel was trained using the appropriate features, i.e. positions 11 and 25, all positions, and LMT’s positions. For the LMT’s positions, we used 9 positions on the V3 amino acid The simplest predictor uses only positions 11 and 25, since these positions have been reported as a basic method for coreceptor usage prediction as described previously. The best performance in this case is 81.3% accuracy, 81.8% specificity, 80.4% sensitivity using the RBF kernel. However, there were many research groups who have used all 35 positions of the V3 amino acid sequences for coreceptor usage prediction. For example, Sing *et al.* used the SVM with the simplest kernel or linear kernel [87]. In this study, using the RBF kernel, we still obtained the maximum performance which increased dramatically reaching a maximum 97.8% accuracy, 98.3% specificity, 96.9% sensitivity. Finally, we also examined the logistic model tree which was used to select the most relevant positions. Using only 9 positions from the V3 amino acid sequences, the SVM classifier obtained the surprisingly significant performance, with accuracy, specificity, and sensitivity of 97.8%, 97.7%, and 97.9%, respectively.

As indicated, the predictive performance reached about 80.0% when using only positions 11 and 25. The distribution of amino acids at these positions presented in a similar fashion for both the R5 and X4 phenotypes (data not shown). Using only these positions seems to be insufficient for classification of the coreceptor usage in the CRF01_AE clade. On the other hand, when all positions of the V3 amino acid sequences were used in the same criterion, we obtained a maximum performance as high as 98.0% on accuracy and specificity. For the optimized case, we found that using only 9 positions, the performance yielded results similar to the maximum case. It should be noted that the 9 positions from the V3 amino acid sequences determined by the LMT can efficiently classify the HIV-1 coreceptor usage for at least the CRF01_AE clade. In Table 4.3, we list the predictive performances obtained from SVM based on the RBF kernel with different features. Intuitively, more features leads to increased accuracy, but this is not necessary true in this case. A number of researches have reported that many features can have irrelevant and redundant information which can reduce the predictive ability of the classifier.

Table 4.2 Predictive performances of SVM based different features with various kernel functions

Kernel functions	**Features**	**Accuracy (%)**	**Specificity (%)**	**Sensitivity (%)**
Linear	11 and 25	74.7	77.3	70.1
	All	96.7	97.7	94.8
	LMT	96.3	96.0	96.9
Polynomial (d=2)	11 and 25	78.4	83.5	69.1
	All	97.1	97.7	95.9
	LMT	97.4	97.2	97.9
Polynomial (d=3)	11 and 25	77.7	82.4	69.1
	All	97.1	97.7	95.9
	LMT	97.4	97.2	97.9
Polynomial (d=4)	11 and 25	76.9	81.3	69.1
	All	97.1	97.7	95.9
	LMT	96.7	96.6	96.9
RBF	11 and 25	81.3	81.8	80.4
	All	97.8	98.3	96.9
	LMT	97.8	97.7	97.9

Table 4.3 Comparison of predictive performances from different features when using SVM based radial basis function kernel

Features	Accuracy (%)	Specificity (%)	Sensitivity (%)
11 and 25	81.3	81.8	80.4
All positions	**97.8**	**98.3**	96.9
LMT's positions	**97.8**	97.7	**97.9**

4.5.2 Comparison of the kernel methods based LMT to other genotypic predictors and computational classification methods

In Table 4.4, we compare the results from our method which include the two main approaches similar to well-known tools and classifiers. First, the genotypic predictors are addressed as the genotypic determinations of HIV-1 coreceptor usage. These approaches focus on using V3 amino acid sequences interpreted according to different genotypic predictors. A Neural Network (NNs) was selected to compare against our method because these methods are well-known approaches for the classification task. The Neural Network Package in RGui was used as the implementation. This package name is nnet [69], which can be downloaded freely from http://cran.r-project.org/web/packages /e1071/index.html. The values of size (number of units in the hidden layer), range (initial random weights), decay (parameter for weight decay) and maxit (maximum number of iterations) were set as 2, 0.1, 5e-4 and 6000, respectively.

We evaluated the predictive performance between the SVM classifier with the LMT and an RBF kernel versus the other methods to classify HIV-1 coreceptor usage based on V3 amino acid sequences. We first used the basic method or 11/25 rules which obtain 76.2% accuracy. When using net charge and combined rules, the accuracy increased to 85.0% and 86.1%, respectively, but the sensitivities decreased to 69.0%. Moreover, the performance using the 11/25 rule decreased the sensitivity dramatically to 33.0%. Since the 11/25 rule was used to classify between R5 and X4 viruses based on subtype B data and the genetic characteristics of subtype B and CRF01_AE are different, the use of this rule to predict the coreceptor usage of CRF01_AE doesn't necessarily follow. As indicated, we also compared the genotypic predictor which is a bioinformatics tool based on Support Vector Machine, called Geno2pheno (G2P) with different false-positive rates (1% to 20%). The best performance reached 97.1% accuracy, 96.6% specificity, and 97.9% sensitivity when using FPR 2.50% which seemed to be slightly better for predicting tropism in the CRF01_AE clade. The performance of PSSM by using X4R5 matrix ($PSSM_{X4R5}$) showed good accuracy, specificity and sensitivity (85.0, 85.8 and 83.5%, respectively) that was similar to using the SINSI matrix ($PSSM_{SINSI}$) (81.7, 86.4 and 73.5%, respectively), but less than G2P at FPR 2.5%. In addition, we found that using the FPR 1.0% for the genotypic predictor obtained 100.0% specificity, but the sensitivity decreased from 97.9% to 69.1%. For the 9 positions determined by the LMT we achieved a predictive accuracy as high as 97.8%, which is better than the other genotypic predictors. Moreover, the performance of the specificity and sensitivity were 97.7% and 97.9%, respectively, which are close to the best performances of the other classifiers.

Table 4.4 Performances of Kernel method vs. the other genotypic predictors.

Methods	**Accuracy (%)**	**Specificity (%)**	**Sensitivity (%)**
11/25 rule	76.2	**100.0**	33.0
Net charge rule	85.0	95.5	66.0
Combined rule	86.1	95.4	69.1
G2P (1%)	89.0	**100.0**	69.1
G2P (2.5%)	97.1	96.6	**97.9**
G2P (5%)	94.5	92.6	**97.9**
G2P (10%)	87.2	81.3	**97.9**
G2P (15%)	82.4	73.9	**97.9**
G2P (20%)	70.3	55.1	**97.9**
PSSM$_{\text{X4R5}}$	85.0	85.8	83.5
PSSM$_{\text{SINSI}}$	81.7	86.4	73.5
NNs	95.6	97.8	91.8
KSVMs based *11 and 25*	81.3	81.8	80.4
KSVMs based All positions	**97.8**	98.3	96.9
KSVMs based *LMT*	**97.8**	97.7	**97.9**

4.5.3 Discussion

In this study we have found some limitations. First, our dataset (177 isolations showing R5 phenotype, and 96 showing X4 phenotype) did not remove sequence identity. Second, part of the validation results, were not derived from testing an independent set. Therefore, in this thesis, the deletion of duplicate sequence and separated independent sets were used to perform our proposed method. The summary of removed high sequence identity dataset is shown in Table 4.5.

Table 4.5 Summary of removed high sequence identity of HIV-1 coreceptor usage

	All	Training	Testing
R5	84	55	29
X4	43	30	13
Total	127	85	42

Our proposed method is used to perform two types of analysis, i.e. perform with the whole dataset (i.e., 127 sequences with 10-fold cross-validation) and separated training (85 sequences) and testing (42 sequences) dataset to validate the efficient our proposed method. We first performed the analysis with the whole dataset for comparing the result from [2]. The new selected positions are P_5 , P_7 , P_{11} , P_{18}, P_{27} , P_{32}. As seen, the positions of P_{13} , P_{14} , P_{19} do not occur in the new dataset. The SVM classifier was then used to predict the new dataset of HIV-1 CRF_AE coreceptor usage with the new informative positions. The result of using the new selected position is 89.0% accuracy. Finally, the training and testing dataset are used to perform the analysis. The new selected positions of the separating data are P_5 , P_7 , P_{18} , P_{27} , P_{32}. This result shows that the positions of P_5 , P_7 , P_{18} , P_{27} , P_{32} are informative to predict HIV-1 coreceptor usage, since these positions are found in both the all and the separated dataset performed on the new dataset. The result of the testing data found a 92.9% accuracy. Although, the resulting accuracy that was found from the separating data was lower than [2], but in practice, the experimental result are more reliable than our old results.

4.6 Conclusion

We have proposed an approach to classify HIV-1 coreceptor usage using the SVM method based on feature selection. Table 4.3 summarizes the performance comparison among different features when classifying HIV-1 coreceptor usage using the radial basis function kernel. The SVM classifier based V3 amino acid yielded 97.7% specificity and 97.9% sensitivity which were considered as a maximum performance. We also added feature selection which was used to identify and remove irrelevant and redundant information to improve the predictive performance. The LMT method was used to make the feature selection. The SVM classifier based LMT method provided a predictive performance which is similar to the maximum performance in all other classifiers. Moreover, our method achieved a sensitivity as high as 97.9% which was better that the SVM classifier based on all positions of V3 amino acid sequences. Obviously, the features which were selected by the LMT method can reduce the dimensionality of the data, so by using the learning SVM method we could obtain a more effective SVM classifier. Finally, we compared our method with the genotypic predictors. Table 4.4 shows that using the 9 positions led to the performance of the SVM method which yielded the 98.0% specificity. From our observations, using these positions could improve the classification of the coreceptor usage for the CRF01_AE clade, the predominant circulating subtype found in Thailand and Southeast Asia. However, further studies were needed to evaluate this prediction in clinical V3 amino acid sequences isolated from HIV-1 infected patients in Thailand and to confirm whether these positions may influence further functional properties of coreceptor usage.