International Journal of Biological, Life and Agricultural Sciences
ISSN: 2415-6612
Vol:3, No:5, 2009

Analysis of Physicochemical Properties on Prediction of R5, X4 and R5X4 HIV-1 Coreceptor Usage

Kai-Ti Hsu, Hui-Ling Huang, Chun-Wei Tung, Yi-Hsiung Chen, and Shinn-Ying Ho

***Abstract*—**Bioinformatics methods for predicting the T cell coreceptor usage from the array of membrane protein of HIV-1 are investigated. In this study, we aim to propose an effective prediction method for dealing with the three-class classification problem of CXCR4 (X4), CCR5 (R5) and CCR5/CXCR4 (R5X4). We made efforts in investigating the coreceptor prediction problem as follows: 1) proposing a feature set of informative physicochemical properties which is cooperated with SVM to achieve high prediction test accuracy of 81.48%, compared with the existing method with accuracy of 70.00%; 2) establishing a large up-to-date data set by increasing the size from 159 to 1225 sequences to verify the proposed prediction method where the mean test accuracy is 88.59%, and 3) analyzing the set of 14 informative physicochemical properties to further understand the characteristics of HIV-1coreceptors.

***Keywords*—**Coreceptor, genetic algorithm, HIV-1, SVM, physicochemical properties, prediction.

I. Introduction

HIV-1 (Human Immunodeficiency Virus Type 1) integrates with T cell after entering the human body that will cause the immune system function to defect [1]. HIV-1 entry is an attractive target for anti-HIV-1 therapy. HIV-1 entry to macrophages and CD4+ T cells is mediated through interaction of the virion envelope glycoproteins (gp120) with the CD4 molecule on the target cells and also with chemokine coreceptors [2]. The past medicines all suppress the mechanism of duplication [3], but the side effects of these medicines are relatively great, and the drug effect is not very apparent [4]. Therefore, in recent years, the new medicines were developed by preventing the synthesis between the virus and immune cells.

HIV-1 tropism is governed in large part by coreceptor choice. The ability of a virus to use CCR5, CXCR4, or both largely dictates the type of CD4-positive cells it can enter [5]. Based on the above ability, HIV-1 variants are classified as CCR5 tropic (R5), CXCR4 tropic (X4), or dual-mixed tropic (R5X4) [6, 7]. The synthesis of HIV-1 and T cell is cooperated by coreceptor, and the main mechanism of synthesis is the connection between the outer membrane protein part of HIV-1 and the receiving part of T cell [4, 5]. The process of synthesis, the coreceptor acts to stabilize the binding side. Some studies [6, 7, 8] aim to predict the coreceptor of T cell using the array of membrane protein of HIV-1.

The existing methods [9, 10] mainly use the array of V3 loop area in the membrane protein of HIV-1 to predict. Using bioinformatics tools based on V3 sequences, the study [6] evaluated the performances of several genotypic tools to predict HIV-1 tropism in non-B subtypes. The study [7] compared with several statistical learning methods for the prediction of HIV-1 coreceptor usage from clonally HIV-1 third hypervariable (V3) loop sequences, and improved their effectiveness on clinical samples. However, the V3 loop sequences [7,11,12] are 34~42 amino acid arrays in length, where most arrays contain 35 amino acids. Moreover, the mutation of HIV-1 is quite rapid, leading to high mutation area of V3 loop [13]. Therefore, it is difficult to predict the coreceptor.

Bioinformatics methods for predicting HIV-1 phenotypes [8, 14] have built training models from the Los Alamos National Laboratory HIV Sequence Database. The prediction of the T cell coreceptors can be regarded as a two- or three-class classification problem. The three classes of coreceptors are CXCR4 (X4), CCR5 (R5) and CCR5/CXCR4 (R5X4). The study [14] first proposed a three-class prediction method using the coupling of domain-level features and amino acid position features to discriminate R5X4 from R5 and X4. In the two-class studies [8, 14], the coreceptor CCR5/CXCR4 is classified into one of the X4 and R5. In this study, we aim to propose an effective method for dealing with the three-class prediction problem.

Using both informative features and an appropriate classifier is essential to design an effective method for prediction of HIV-1 coreceptor usage. Besides pursuit of high prediction accuracy, mining novel and interpretable features

K.-T. Hsu is with Institute of Bioinformatics and Systems Biology, the National Chiao Tung University, Hsinchu, Taiwan (e-mail: cafehe.bi96g@g2.nctu.edu.tw).

H.-L. Huang is with Institute of Bioinformatics and Systems Biology, Department of Biological Science and Technology, the National Chiao Tung University, Hsinchu, Taiwan (e-mail: hlhuang@mail.nctu.edu.tw).

C.-W. Tung is with Institute of Bioinformatics and Systems Biology, the National Chiao Tung University, Hsinchu, Taiwan (e-mail: cwtung@livemail.tw).

Y.-H. Chen is with Institute of Bioinformatics and Systems Biology, the National Chiao Tung University, Hsinchu, Taiwan (e-mail: blair.bi94g@g2.nctu.edu.tw).

S.-Y. Ho is with the Institute of Bioinformatics and Systems Biology, and Department of Biological Science and Technology, National Chiao Tung University, Hsinchu, Taiwan (corresponding author to provide phone: 886-3-571-2121, ext: 56909; e-mail: syho @mail.nctu.edu.tw).

are also taken into account in this study. Physicochemical properties extracted from protein sequences were utilized as effective features in recent years. ProLoc [15] is a support vector machine (SVM) based classifier with automatic selection from a large set of physicochemical composition features to predict protein subnuclear localization. The method POPI used physicochemical properties as effective features to predict peptide immunogenicity [16]. The prediction method UbiPred [17] mined informative physicochemical properties from protein sequences to identify promising ubiquitylation sites.

In this study, we made efforts in the three aspects: 1) proposed a feature set of informative physicochemical properties which is cooperated with SVM to achieve high prediction test accuracy of 81.48%, compared with the existing method [14] with accuracy of 70.00%; 2) established a larger up-to-date data set by increasing the size from 159 to 1225 sequences to verify the proposed prediction method where the mean test accuracy is 88.59%, and 3) analyzed the set of 14 informative physicochemical properties to further understand the characteristics of HIV-1coreceptors.

II. Materials

A. Datasets

In this study, we established three data sets with various sizes for evaluating the proposed methods and comparing with the existing methods Data1225, Data139 and Data2class.

1) Data1225. All the sequence data of HIV-1 V3 loop were downloaded from the Los Alamos National Laboratory HIV Sequence Database (http://www.hiv.lanl.gov/) [18]. Sequences containing characters that excluding 20 amino acids were removed from the dataset. The remaining samples have 2,940 sequences, including 2,235 R5-utilizing sequences, 383 X4-utilizing sequences, and 322 R5X4-utilizing sequences.

In order to obtain more general and impartial model, the sequences were filtered in two aspects. First, we eliminated the sequences of redundancy. For example, the sequence, CTRPSNNTRTSITIGPGQVWYRTGDIIGNIRKAYC, appears in 24 protein sequences with different accession numbers. To avoid overfitting or prejudiced model, 1,671 redundant sequences were eliminated. Secondly, we removed sequences which come into conflict with each other. That means if a sequence appears at two or more categories of coreceptor usage, we removed the sequences from the dataset. For example, the sequence, CTRPYNNTRQSTHIGPGRAYYTTKIIGDIRQAHC, exists at both R5 and X4 classes of coreceptor usage. After discarding the sequences of conflict, we obtain 1,225 sequences (931 R5, 164 X4, and 130 R5X4).

2) Data2class. The data set obtained from [7] has 1,110 sequences. We deleted the redundant and conflict sequences. Consequently, the resultant data set Data2class has 507 sequences with only two categories, R5-only and X4-capable. The R5-only category has R5-utilizing sequences only, and the X4-capable category contains X4-utilizing and/or R5X4-utilizing sequences [7].

3) Data139. For comparing the prediction method [14], which is the first method to predict three categories of HIV-1 coreceptor usage directly, we established a data set Data139. The used data set in [14], named Data157, has 157 sequences of three categories. Since the sequences of Data157 are not available, we collected the sequences according to the accession numbers provided by the work [14]. Due to the upgrade of the HIV database probably, some accession numbers cannot find the corresponding sequence.

Therefore, the data set Data139 has 72 R5, 40 X4, and 27 R5X4 sequences. The statistic of the four datasets is shown in Table I.

B. Physicochemical Properties

Physicochemical properties are the most intuitive feature for biochemical reactions and are extensively applied in bioinformatics studies. The amino acid indices (AAindex) database collects many published indices representing physicochemical properties of amino acids. For each physicochemical property, there is a set of 20 numerical values for amino acids. Currently, 544 physicochemical properties can be retrieved from the AAindex database of version 9.0 [19]. After removing physicochemical properties having the value 'NA' in the amino acid indices, 531 physicochemical properties are obtained for the following studies. In contrast to the residue-based encoding methods of amino acid identity and evolutionary information, there are 531 mean values used to represent a sample [15, 16]. If *m* out of 531 informative physicochemical properties are selected and are used in support vector machine (SVM), *m* mean values are used to represent a sample.

TABLE I
STATISTIC OF THE DATASETS

Datasets	Number of R5-usage	Number of X4-usage	Number of R5X4-usage	Number of Total Sequences
Data2class	423	84	0 [a]	507
Data157	82	36	39	157
Data139	72	40	27	139
Data1225	931	164	130	1225

The statistic of the datasets includes the numbers of three categories and total numbers of sequences.
[a]Data2class combined the X4-usage and R5X4-usage to X4-capable, so that, the number of X4-usage includes sequences of two classes.

III. Methods

We propose a novel SVM-based method (named SVM-PCP) using physicochemical properties for predicting HIV-1 coreceptors. The identification of an effective feature set of physicochemical properties is mainly derived by using an inheritable bi-objective genetic algorithm (IBCGA) [20]. The IBCGA mines informative physicochemical properties and tune parameter settings of SVM simultaneously while maximizing 10-fold cross validation (10-CV) accuracy. The selected m=14 physicochemical properties and the designed SVM are used to implement the computational system HIV1-Cor for prediction of R5, X4 and R5X4 coreceptors.

A. Support Vector Machine

Support vector machine (SVM) is a learning model dealing with binary classification problems. SVM constructs a binary classifier by finding a hyperplane to separate two classes with a maximal distance between margins of two classes consisting of support vectors. In order to make linear separation of samples easier, SVM uses one of various kernel functions to transform the samples into a high-dimensional search space. In this work, the commonly-used radial basis function is applied to nonlinearly transform the feature space, defined as follows:

$$K(x_i, x_j) = \exp(-\gamma \left\| x_i - x_j \right\|), \gamma > 0 \quad (1)$$

The kernel parameter γ determines how the samples are transformed into a high-dimensional search space. The cost parameter C>0 of SVM adjusts the penalty of total error. These two parameters C and γ must be tuned to get the best prediction performance.

For multi-class classification problems, 'one-against-one' strategy is applied to transform the multi-class problem into several binary classification problems. Given h classes, there are $h(h-1)/2$ classifiers constructed and each one trains the samples from two classes. A voting strategy is applied to give a final prediction for test samples. In this study, h=3 and the used SVM is obtained from LIBSVM package version 2.81 [21].

B. Inheritable Bi-objective Genetic Algorithm

Selecting a minimal number of informative features while maximizing prediction accuracy is a bi-objective 0/1 combinatorial optimization problem. An efficient inheritable bi-objective genetic algorithm [22] is utilized to solve this optimization problem. IBCGA consists of an intelligent genetic algorithm [23] with an inheritable mechanism. The intelligent genetic algorithm uses a divide-and-conquer strategy and an orthogonal array crossover operation to efficiently solve large-scale parameter optimization problems. In this study, the intelligent genetic algorithm can efficiently explore and exploit the search space of C(n, r) to identify r out of n physicochemical properties. IBCGA can efficiently search the space of C(n, $r \pm 1$) by inheriting a good solution in the space of C(n, r) [22]. Therefore, IBCGA can economically obtain a complete set of high-quality solutions in a single run where r is specified in an interesting range such as [5, 50].

The proposed chromosome encoding scheme of IBCGA consists of both binary genes for feature selection and parametric genes for tuning SVM parameters, where the gene and chromosome are commonly-used terms of genetic algorithm (GA), named GA-gene and GA-chromosome for discrimination in this paper. The GA-chromosome consists of n=531 binary GA-genes b_i for selecting informative properties and two 4-bit GA-genes for tuning the parameters C and γ of SVM. If b_i=0, the i^{th} property is excluded from the SVM classifier; otherwise, the i^{th} property is included. This encoding method maps the 16 values of γ and C into $\{2^{-7}, 2^{-6}\ldots, 2^{8}\}$.

The feature vector for training the SVM classifier is obtained from decoding a GA-chromosome using the following steps. Consider a given HIV-1 sequence. At first, the index vectors for all selected physicochemical properties are constructed from AAindex for each amino acid. Feature vector of a peptide consists of the selected features whose values are obtained by averaging the values in their corresponding index vectors. Finally, all values of the feature vectors are normalized into [-1, 1] for applying SVM.

Fitness function is the only guide for IBCGA to obtain desirable solutions. The fitness function of IBCGA is the 10-CV overall accuracy.

IBCGA with the fitness function f(X) can simultaneously obtain a set of solutions, X_r, where r=r_{start}, r_{start}+1, …, r_{end} in a single run. The algorithm of IBCGA with the given values r_{start} and r_{end} is described as follows:

Step 1) (Initiation) Randomly generate an initial population of N_{pop} individuals. All the n binary GA-genes have r 1's and n-r 0's where $r = r_{start}$.

Step 2) (Evaluation) Evaluate the fitness values of all individuals using f(X).

Step 3) (Selection) Use the traditional tournament selection that selects the winner from two randomly selected individuals to form a mating pool.

Step 4) (Crossover) Select $p_c \cdot N_{pop}$ parents from the mating pool to perform orthogonal array crossover on the selected pairs of parents where p_c is the crossover probability.

Step 5) (Mutation) Apply the swap mutation operator to the randomly selected $p_m \cdot N_{pop}$ individuals in the new population where p_m is the mutation probability. To prevent the best fitness value from deteriorating, mutation is not applied to the best individual.

Step 6) (Termination test) If the stopping condition for obtaining the solution X_r is satisfied, output the best individual as X_r. Otherwise, go to Step 2). In this study, the stopping condition is to perform 40 generations.

Step 7) (Inheritance) If $r < r_{end}$, randomly change one bit in the binary GA-genes for each individual from 0 to 1; increase the number r by one, and go to Step 2). Otherwise, stop the algorithm.

C. Prediction Method SVM-PCP

The selected m physicochemical properties and the associated parameter set of SVM by using IBCGA are used to implement the computational system HIV1-Cor and analyze the physicochemical properties to further understand the coreceptors. Since the IBCGA is a non-deterministic method, it should make more effort to identify an efficient and robust feature set of informative physicochemical properties. First, we prepare the K independent data sets where each set is used as the training data set of 10-CV. Secondly, IBCGA is performed R independent runs for each of K data sets. In this study, $K = 10$ and $R = 20$. Therefore, there are total K x R = 200 runs of IBCGA. There are total 200 sets of m physicochemical properties. Thirdly, we calculate and record the selection frequencies $F()$ of the selected physicochemical properties from the solutions of 200 independent runs on training datasets. Fourthly, we calculate score S_r (r = 1, …, KxR) for each solution as follows:

$$S_r = (\sum_{i=1}^{m} F(P_i))/m \quad (2)$$

where $F(P_i)$ denotes the frequencies of the physicochemical property P_i, m is the number of the selected feature in the r-th dependent run. Finally, choose the set of selected physicochemical properties with a maximal value of S_r.

The HIV1-Cor system flowchart using the prediction method SVM-PCP is shown in Fig. 1. SVM-PCP will automatically determine a set of informative physicochemical properties and an SVM-model for prediction HIV-1 coreceptor levels. The frequencies of the selected physicochemical properties from the solutions of KxR = 200 independent runs on Data1225 dataset are shown in Fig. 2.

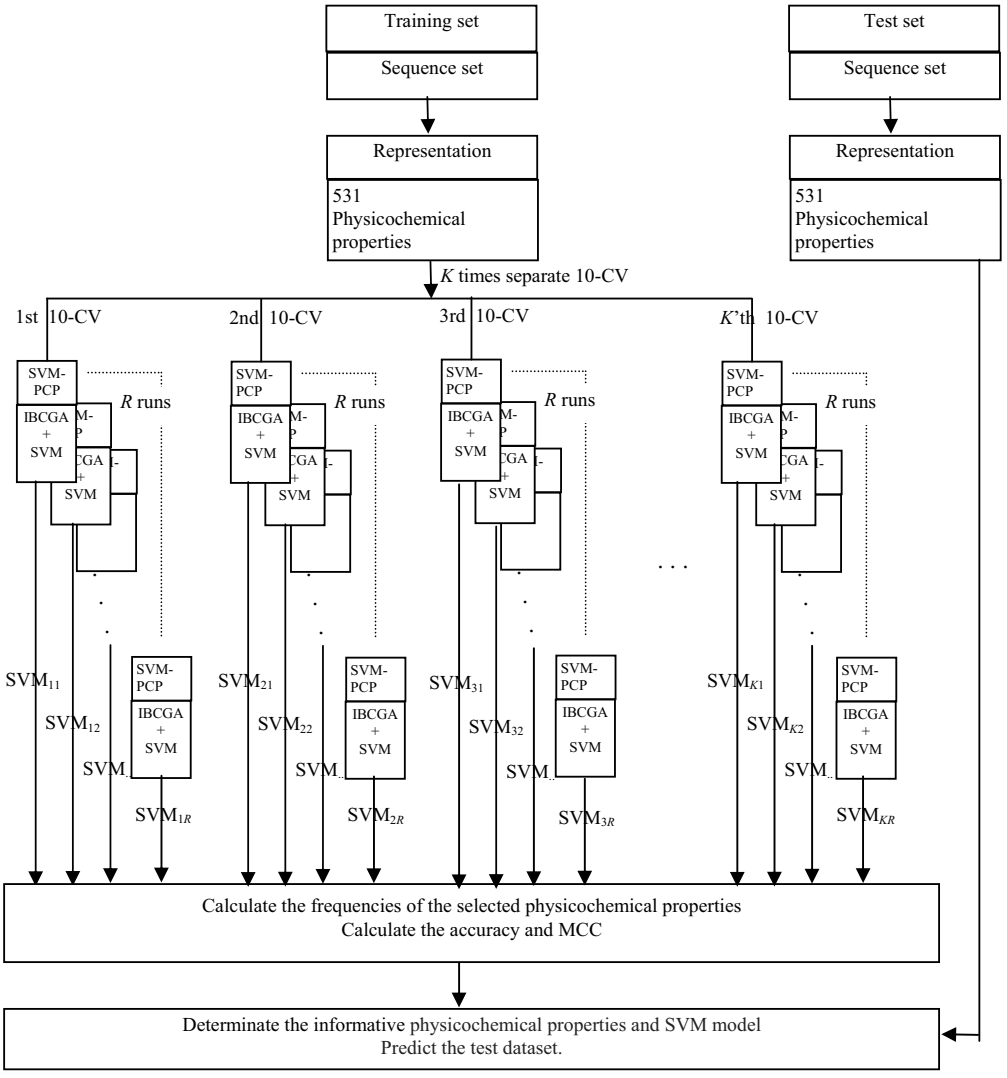


Fig.1. The system flowchart of the prediction method SVM-PCP

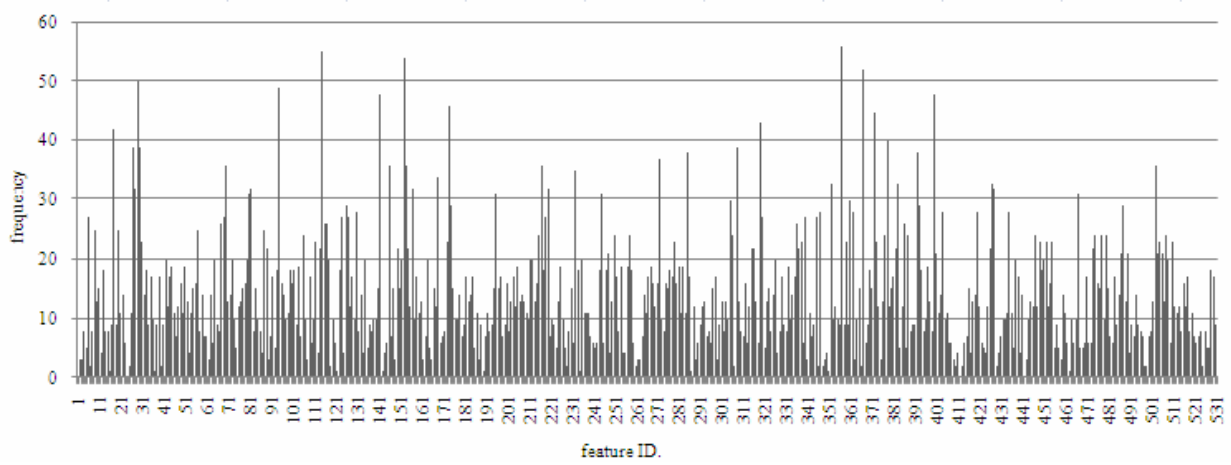


Fig.2. Frequencies of the selected physicochemical properties from the solutions of K x R = 200 independent runs on Data1225 dataset.

IV. Results

A. Environments

Several alternatives of prediction based on statistical learning methods have been developed [24], including linear regression [25], artificial neural networks [26], decision trees [26], SVM [27], position-specific scoring matrices (PSSM) [28] and mixtures of localized rules [29]. The best of those methods is SVM [7]. As a result, we utilized SVM in our method. We incorporate LIBSVM into SVM-PCP as classifier and use the default values of SVM parameters were used.

SVM-PCP is performed to select the physicochemical properties to be the features of SVM, and the parameters of IBCGA are r_{start} = 5, r_{end} = 45, and default values of other parameters.

B. Training and independent test sets

We divided the datasets into training and independent test sets. Data2class is randomly separated to five parts, four parts for training and one part for independent test. The training/test data of Data2class is called Set2class. Set157 is obtained from Data157 and we show the number of the training and test data come from the [14]. Set139 is spilt from Data139 using the ratio of the training/test data of Set157 (about 4:1 with training and test data). Set1225 and Set3fold are obtained from the Data1225, but used different ratios of training and test data 4:1 and 2:1, respectively, shown in Table II.

TABLE II
Training/test Sets of Datasets

Datasets	Training Set	Test Set	Total Number
Set2class	405	102	507
Set157	127	30	157
Set139	112	27	139
Set1225	991	234	1225
Set3fold	817	408	1225

C. Results of two categories of coreceptor

In previous studies, the prediction of HIV-1 coreceptor usage is concentrated on two categories of coreceptor, R5 and X4, and the class R5X4-using coreceptors usually were concluded to X4 classes [7]. Set2class is used to compare the study, which predicting the HIV-1 coreceptor usage by two classes [7]. Because of the feature sources and learning method, Set2class was named SVM-PCP to distribute the SVM method of previous study [7]. The result was shown in Table III. The 11/25 rule has a mean sensitivity of 59.5% in detecting X4-capable variants and a mean specificity of 92.5% [7]. In order to analyzing the ability of prediction with 11/25 rule, the ROC curve was used to estimate the sensitivity and specificity of detecting X4-capable variants and we also computed the areas under the ROC curve (AUC) to analyze [7]. Considering the 11/25 rule, we fixing the specificity at 92.5% to analyze the sensitivity and compared the AUC. Result of SVM-PCP at sensitivity is obvious outstanding than previous studies and the AUC is also better than others.

We tested the training model using an independent dataset including 102 sequences, and the average number of correct prediction is 91.84 sequences with accuracy 90.22%. The scenario means that no significant overfitting was occurred.

TABLE III
Results at Fixed Specificity

Method	Sensitivity (%)	AUC
11/25 rule	59.5	*
PSSM	71.9	0.90
[a]SVMbinary	76.4	0.91
SVM-PCP	88.9	0.94

The results of sensitivity are obtained by fixed the specificity at 92.5%. 11/25 rule is the earliest method to predict the co-receptor usage. Although, it has high specificity, the sensitivity of 11/25 rule is lower. 11/25 rule just only has one result so that no ROC curve is available. PSSM and SVMbinary improve the result of prediction but the best result is SVM-PCP.

D. Results of three categories of coreceptor

Prediction of coreceptor usage with three categories is a novel attempt that proposed in 2008 [14]. As a result, we also classify the data into the three classes and predict the coreceptor usage. Data139 is obtained from the Data157, so we can compare their performance directly. We got the K = 10 training/test sets randomly and each set had R = 20 rounds of K x R = 200 experiment. The statistic of training models of Set157 and Set139 were shown in Table IV. Hence Set157 only shows the best result, we compare the best result and also present the average value. For independent test sets, we not only show the compared of best result and average results, but also show the result of sequences in Table V. Comparing the best data, Set139 had higher accuracy and MCC at all the classes. Matthew's correlation coefficient (MCC_i) for the i^{th} coreceptor class, i = 1, 2, 3, and the best accuracy and averaged accuracies for all classes:

$$MCC_i = \frac{TP_i \times TN_i - FP_i \times FN_i}{\sqrt{(TP_i + FN_i) \times (TP_i + FP_i) \times (TN_i + FP_i) \times (TN_i + FN_i)}} \quad (3)$$

where TP_i, TN_i, FP_i and FN_i are the number of true positive, true negative, false positive and false negative, respectively.

TABLE IV
STATISTIC OF TRAINING DATA OF SET157 AND SET139

	Statistics	Set157 [14] (best)	Set139 (best)	Set139(average)
R5	Accuracy (%)	75.00	99.00	98.33 ± 1.53
	MCC	0.58	0.84	0.82 ± 0.03
X4	Accuracy (%)	79.31	81.50	81.05 ± 4.56
	MCC	0.66	0.85	0.83 ± 0.04
R5X4	Accuracy (%)	40.00	72.73	69.16 ± 7.54
	MCC	0.18	0.75	0.72 ± 0.05
Overall Accuracy (%)		67.72	89.15	87.97 ± 1.60
Mean of Accuracies (%)		64.77	84.41	82.85 ± 2.35

E. Results of different ratios of training and test datasets

To compare the same dataset with the different ratio of training/test sets, Set1225 and Set3fold were used. We also randomly spilt K = 10 training/test sets and each set has R = 20 rounds of experiment. Results of training sets were shown in Table VI and the independent test results were shown in Table VII. As shown in Table VII, larger size of training data has higher accuracy of test.

TABLE V
STATISTIC OF INDEPENDENT TEST OF SET157 AND SET139

	Statistics	Set157 (best)	Set139 (best)	Set139(average)
R5	Accuracy (%)	78.57 (11/14) [a]	91.67 (11/12)	90.21 ± 8.90
	MCC	0.47	0.71	0.64 ± 0.14
X4	Accuracy (%)	70.00 (7/10)	80.00 (8/10)	68.55 ± 15.90
	MCC	0.47	0.76	0.63 ± 0.14
R5X4	Accuracy (%)	50.00 (3/6)	60.00 (3/5)	20.20 ± 17.13
	MCC	0.38	0.61	0.10 ± 0.21
Overall Accuracy (%)		70.00	81.48	69.22 ± 7.48
Mean of Accuracies (%)		66.19	77.22	59.65 ± 7.95

[a] (Number of correct prediction sequences / Total number of sequences)

The summarized test accuracies are shown in Table VIII. The proposed method SVM-PCP can achieve high prediction test accuracies with 69.2222±7.4759, 88.5954±1.3985 and 87.6185±1.0703 for Set139, Set1225 and Set3fold, respectively.

TABLE VI
STATISTIC OF TRAINING DATA OF SET1225 AND SET3FOLD

	Statistics	Set1225 (best)	Set1225(average)	Set3fold (best)	Set3fold(average)
R5	Accuracy (%)	99.20	99.01 ± 0.34	98.23	99.02 ± 0.35
	MCC	0.86	0.82 ± 0.04	0.85	0.8057 ± 0.02
X4	Accuracy (%)	82.71	82.50 ± 2.18	85.32	81.17 ± 2.97
	MCC	0.83	0.83 ± 0.02	0.86	0.83 ± 0.02
R5X4	Accuracy (%)	58.09	51.38 ± 7.97	64.37	48.04 ± 5.36
	MCC	0.66	0.62 ± 0.05	0.69	0.59 ± 0.045
Overall Accuracy (%)		92.63	91.75 ± 1.01	92.90	91.21 ± 0.64
Mean of Accuracies (%)		80.00	77.63 ± 2.76	82.64	76.08 ± 1.88

TABLE VII
STATISTIC OF INDEPENDENT TEST OF SET1225 AND SET3FOLD

	Statistics	Set1225 (best)	Set1225(average)	Set3fold (best)	Set3fold(average)
R5	Accuracy (%)	97.19(173/178) [a]	97.44 ± 1.20	97.10(301/310)	97.49 ± 1.01
	MCC	0.85	0.77 ± 0.05	0.79	0.75 ± 0.04
X4	Accuracy (%)	87.10(27/31)	79.10 ± 5.59	81.82(45/55)	73.55 ± 5.90
	MCC	0.76	0.76 ± 0.04	0.79	0.73 ± 0.04
R5X4	Accuracy (%)	48.00(12/25)	37.41 ± 8.19	44.19(19/43)	34.45 ± 5.83
	MCC	0.56	0.45 ± 0.09	0.50	0.39 ± 0.05
Overall Accuracy (%)		90.60	88.60 ± 1.40	89.46	87.62 ± 1.07
Mean of Accuracies (%)		77.43	71.32 ± 2.91	74.37	68.50 ± 2.58

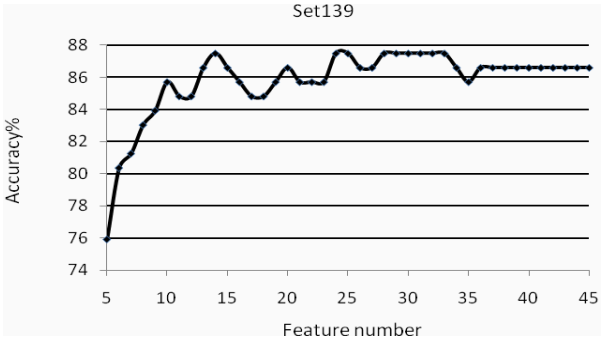
[a] (Number of correct prediction sequences / Total number of sequences)

TABLE VIII
COMPARING OF ACCURACY AND SEQUENCES STATISTIC

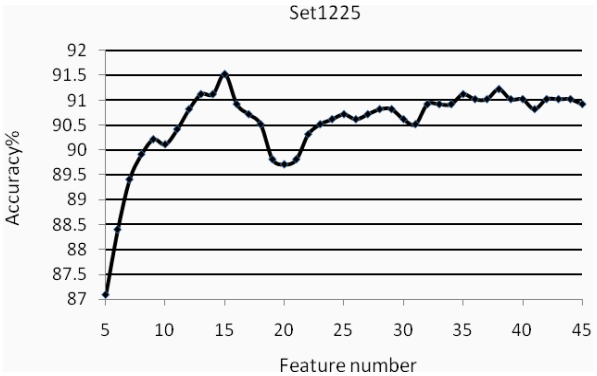
	Statistics	Set139 (average)	Set1225 (average)	Set3fold (average)
R5	Accuracy (%)	90.2083 ± 8.8979	97.4382 ± 1.2016	97.4895 ± 1.0069
	Correct number(n)	11 ± 1	175 ± 3	302 ± 3
	Total number(n)	12	178	310
X4	Accuracy (%)	68.5500 ± 15.8953	79.0968 ± 5.5944	73.5502 ± 5.9010
	Correct number(n)	7 ± 3	25 ± 2	41 ± 3
	Total number(n)	10	31	55
R5X4	Accuracy (%)	20.2000 ± 17.1297	37.4133 ± 8.1919	34.4504 ± 5.8295
	Correct number(n)	1 ± 1	9 ± 2	15 ± 3
	Total number(n)	5	25	43
Overall Accuracy (%)		69.2222 ± 7.4759	88.5954 ± 1.3985	87.6185 ± 1.0703
Correct number(n)		19 ± 2	209 ± 4	358 ± 4
Total number(n)		27	234	408

TABLE IX
THE 14 INFORMATIVE PHYSICOCHEMICAL PROPERTIES DETERMINED BY HIV-1COR ON SET3FOLD

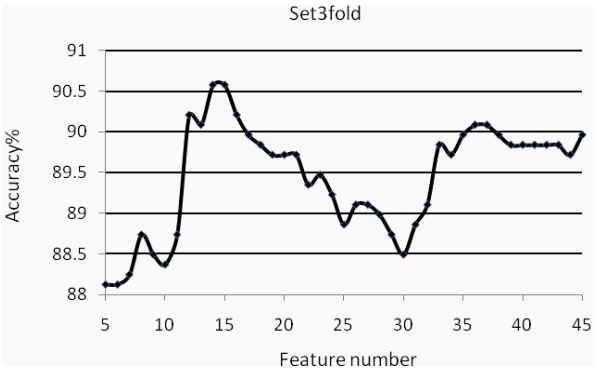
Feature ID.	AAindex identity	Description
30	CHAM830107	A parameter of charge transfer capability (Charton-Charton, 1983)
43	CHOP780206	Normalized frequency of N-terminal non helical region (Chou-Fasman, 1978b)
70	EISD860102	Atom-based hydrophobic moment (Eisenberg-McLachlan, 1986)
95	FINA910104	Helix termination parameter at posision j+1 (Finkelstein et al., 1991)
142	KARP850101	Flexibility parameter for no rigid neighbors (Karplus-Schulz, 1985)
205	NAKH920104	AA composition of EXT2 of single-spanning proteins (Nakashima-Nishikawa, 1992)
281	QIAN880124	Weights for beta-sheet at the window position of 4 (Qian-Sejnowski, 1988)
320	RADA880107	Energy transfer from out to in(95%buried) (Radzicka-Wolfenden, 1988)
335	RICJ880114	Relative preference value at C1 (Richardson-Richardson, 1988)
360	SNEP660102	Principal component II (Sneath, 1966)
386	WERD780103	Free energy change of alpha(Ri) to alpha(Rh) (Wertz-Scheraga, 1978)
392	WOLS870103	Principal property value z3 (Wold et al., 1987)
475	TSAJ990102	Volumes not including the crystallographic waters using the ProtOr (Tsai et al., 1999)
479	WILM950102	Hydrophobicity coefficient in RP-HPLC, C8 with 0.1%TFA/MeCN/H2O (Wilce et al. 1995)



(a)



(b)



(c)

Fig. 3 IBCGA is performed to mine informative physicochemical properties using the whole dataset (a) Set139, (b) Set1225, and (c) Set3fold

F. Selected a small set of physicochemical properties

IBCGA is utilized to mine informative physicochemical properties using the whole dataset [18]. The best results of 200 runs shown in Fig. 3 reveal that the best numbers of selected features is m = 14, 15, 14 where the parameter settings (C, γ) of the SVM classifier are (4, 1.00), (2, 2.00), and (4, 2.00), and the 10-CV accuracies are 91.07%, 91.52%, and 90.58%, for the data sets Set139, Set1225 and Set3fold, respectively.

The feature set with m = 14 having the highest appearance frequency of properties in the 200 feature sets on Set3fold is given in Table IX.

V. CONCLUSIONS

We have proposed a novel method SVM-PCP using physicochemical properties for predicting HIV-1 coreceptor with the three-class prediction problem. The three classes of coreceptors are X4, R5 and R5X4. We have established three data sets with different sizes for evaluating the proposed methods. The IBCGA mines informative physicochemical properties and tune parameter settings of SVM simultaneously while maximizing 10-CV accuracy. We have calculated the frequency statistics of the selected physicochemical properties from the solutions of the independent runs. Determinate the informative physicochemical properties and SVM-model can be predicted the HIV-1 coreceptor. The SVM-PCP can achieve high prediction test accuracy. The results were shown better the existing method. The 14 informative physicochemical properties determined by HIV-1Cor on Set3fold are given. The most important feature work is to further analyze the informative physicochemical properties.

REFERENCES

[1] D. Unutmaz, "T cell signaling mechanisms that regulate HIV-1 infection," *Immunologic Research*, vol. 23, no. 2-3, pp. 167-177, June 2001.

[2] D. C. Chan,. D. Fass, J. M. Berger, "Core Structure of gp41 from the HIV Envelope Glycoprotein," *Cell*, vol. 89, pp. 263–273, April 18, 1997.

[3] M. L. Greenberg, S. F. Lacey, C.-Ho Chen, Dani P. Bolognesi and Kent J. Weinhold, "T cell signaling mechanisms that regulate HIV-1 infection," *Springer Seminars in Immunopatholog* , vol. 18, no. 3, pp. 355-369, Sep. 1997.

[4] M. S. Hirsch, T.-C. Chou, V. A. Johnson, M. A. Barlow, D. P. Merrill, "Three-drug synergistic inhibition of HIV-1 replication in vitro by zidovudine, recombinant soluble CD4, and recombinant interferon-alpha A," *Journal of Infectious Diseases*, Health, 1990.

[5] T. L. Hoffman, C. C. Labranche, W. Zhang, G. Canziani, J. Robinson, I. Chaiken, J. A. Hoxie, and R. W. Doms, "Stable exposure of the coreceptor-binding site in a CD4-independent HIV-1 envelope protein," *Proc. Natl. Acad. Sci. USA*, Medical Science, Vol. 96, May 1999, pp. 6359–6364.

[6] C. Garrido, V. Roulet, N. Chueca, E. Poveda, A. Aguilera, K. Skrabal, N. Zahonero, S. Carlos, F. García, J. L. Faudon, V. Soriano, and C. d. Mendoza, "Evaluation of Eight Different Bioinformatics Tools To Predict Viral Tropism in Different Human Immunodeficiency Virus Type 1 Subtypes," *Journal of Clinical Microbiology*, vol. 46, no. 3, pp. 887–891, Mar. 2008.

[7] T. Sing, A. J Low, N. Beerenwinkel, O. Sander, P. Cheung, F. S Domingues, J. Büch, M. Dämer, R. Kaiser, T. Lengauer and P R. Harrigan "Predicting HIV coreceptor usage on the basis of genetic and clinical covariates," *Antiviral therapy*, vol. 12, pp. 1097–1106, 2007

[8] S. Boisvert, M. Marchand, F. Laviolette and J. Corbei, "HIV-1 coreceptor usage prediction without multiple alignments: an application of string kernels," *Retrovirology*, vol.5, no.110, doi:10.1186/1742-4690-5-110, 2008.

[9] Cormier, E., and T. Dragic, "The crown and stem of the V3 loop play distinct roles in human immunodeficiency virus type 1 envelope glycoprotein interactions with the CCR5 coreceptor," *J. Virol.*, vol.76, pp.8953–8957, 2002.

[10] O. Sander, T. Sing, I. Sommer, A. Low, P. Cheung, P. Harrigan, T. Lengauer,and F. Domingues, "Structural descriptors of gp120 V3 loop for the prediction of HIV-1 coreceptor usage," *PLoS Comput. Biol*. 3:e58, 2007.

[11] W. Resch, N. Hoffman, and R. Swanstrom, "Improved success of phenotype prediction of the HIV type 1 from envelope variable loop 3 sequence using neural networks," *Virology,* vol. 288, pp.51–62, 2001.

[12] C. Pastore, R. Nedellec, A. Ramos, S. Pontow, L. Ratner, and D. Mosier, "HIV type 1 coreceptor switching: V1/V2 gain-of-fitness mutations compensate for V3 loss-of-fitness mutations," *J. Virol.*, vol. 80, pp. 750–758, 2006.

[13] J. Weber, "The pathogenesis of HIV-1 infection," *British Medical Bulletin,* vol. 58, pp.61-72, 2001.

[14] S. L. Lamers, M. Salemi, M. S. McGrath, and G. B. Fogel, "Prediction of R5, X4, and R5X4 HIV-1 Coreceptor Usage with Evolved Neural Networks," in *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 5, No. 2, pp. 291-300, April-June 2008.

[15] W.-L. Huang, C.-W. Tung, H.-L. Huang, S.-F. Hwang, S.-Y. Ho, "ProLoc: Prediction of protein subnuclear localization using SVM with automatic selection from physicochemical composition features," *BioSystems,* vol. 90, pp. 573–581, 2007.

[16] Chun-Wei Tung and Shinn-Ying Ho, "POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties," *Bioinformatics*, vol. 23, no. 8, pp. 942–949, 2007.

[17] Chun-Wei Tung and Shinn-Ying Ho, "Computational identification of ubiquitylation sites from protein sequences," *BMC Bioinformatics*, vol. 9:310, July 2008.

[18] Los Alamos National Laboratory HIV Sequence Database, http://www.hiv.lanl.gov/**.**

[19] S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, M. Kanehisa, "AAindex: amino acid index database," progress report 2008**.** *Nucleic Acids Res* 2008, 36(Database issue):D202-205.

[20] JR Quinlan. *C4.5: programs for machine learning*. In. San Mateo, CA: Morgan Kaufmann. 1993.

[21] C. C. Chang, and, C. J. Lin (2001) *LIBSVM: a library for support vector machines*. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[22] S. Y. Ho, J. H. Chen, and M. H. Huang.,"Inheritable genetic algorithm for bi-objective 0/1 combinatorial optimization problems and its applications," *IEEE Trans. Syst. Man Cybern. Part B-Cybern.*, vol. 34, pp. 609-620, 2004a.

[23] S. Y. Ho, L. S. Shu, and J. H. Chen," Intelligent evolutionary algorithms for large parameter optimization problems," *IEEE Trans. Evol. Comput.*, vol. 8, pp. 522-541, 2004b.

[24] M. A. Jensen and A. B. van 't Wout, "Predicting HIV-1 coreceptor usage with sequence analysis," *IDS Rev*, vol. 5, 2003, pp. 104-112.

[25] D. R. Briggs, D. L. Tuttle, J. W. Sleasman, M. M. Goodenow, "Envelope V3 amino acid sequence predicts HIV-1 phenotype (coreceptor usage and tropism for macrophages," *AIDS*, vol. 14, 2000, pp. 2937-2939.

[26] W. Resch, N. Hoffman, R. Swanstrom, "Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks," *Virology*, vol. 288, pp. 51-62, 2001

[27] S. Pillai, B Good, D. Richman, J. Corbeil, "A new perspective on V3 phenotype prediction," *AIDS Res Hum Retroviruses*, vol. 19, pp. 145-149, 2003.

[28] M. A. Jensen, *et al*., "Improved coreceptor usage prediction and genotypic monitoring of R5-toX4 transition by motif analysis of human immunodeficiency virus type 1 *env* V3 loop sequences," *J Virol*, vol. 77, pp. 13376-13388, 2003.

[29] T. Sing, N. Beerenwinkel, T. Lengauer, "Learning mixtures of localized rules by maximizing the area under the ROC curve," *Proceedings of 1st International Workshop on ROC Analysis in Artificial Intelligence*, 22 August 2004, pp. 89-96