Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

OXFORD

Subject Section

Genetic source completeness of HIV-1 circulating recombinant forms (CRFs) predicted by multi-label learning

Runbin Tang [1,2], Zuguo Yu [1,3*], Yuanlin Ma [1], Yaoqun Wu [1], Yi-Ping Phoebe Chen [4], Limsoon Wong [5] and Jinyan Li [2,*]

[1]Key Laboratory of Intelligent Computing and Information Processing of Ministry of Education and Hunan Key Laboratory for Computation and Simulation in Science and Engineering, Xiangtan University, Hunan 411105, China; [2]Advanced Analytics Institute, University of Technology Sydney, NSW 2007, Australia; [3]School of Electrical Engineering and Computer Science, Queensland University of Technology, QLD 4001, Australia; [4]Department of Computer Science and Information Technology, La Trobe University, Victoria 3086, Australia; and [5]School of Computing, National University of Singapore, Singapore 117417.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Infection with strains of different subtypes and the subsequent crossover reading between the two strands of genomic RNAs by host cells' reverse transcriptase are the main causes of the vast HIV-1 sequence diversity. Such inter-subtype genomic recombinants can become circulating recombinant forms (CRFs) after widespread transmissions in a population. Complete prediction of all the subtype sources of a CRF strain is a complicated machine learning problem. It is also difficult to understand whether a strain is an emerging new subtype and if so, how to accurately identify the new components of the genetic source.
Results: We introduce a multi-label learning algorithm for the complete prediction of multiple sources of a CRF sequence as well as the prediction of its chronological number. The prediction is strengthened by a voting of various multi-label learning methods to avoid biased decisions. In our steps, frequency and position features of the sequences are both extracted to capture signature patterns of pure subtypes and CRFs. The method was applied to 7185 HIV-1 sequences, comprising 5530 pure subtype sequences and 1655 CRF sequences. Results have demonstrated that the method can achieve very high accuracy (reaching 99%) in the prediction of the complete set of labels of HIV-1 recombinant forms. A few wrong predictions are actually incomplete predictions, very close to the complete set of genuine labels.
Availability: https://github.com/Runbin-tang/The-source-of-HIV-CRFs-prediction
Contact: yuzuguo@aliyun.com;jinyan.li@uts.edu.au
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genetic sources and mutation characteristics of human immunodeficiency viruses (HIV) have been extensively studied to understand their origins and evolution patterns for better drug design and disease treatment (Moutouh *et al.*, 1996; Rambaut *et al.*, 2004; Zhang *et al.*, 2010). HIV genome is composed of two strands of RNAs (of about 9.7 thousand bases each) packaged inside a cone-shaped capsid (Rajarapu, 2014). Although the two strands of viral RNAs are very similar, millions of variants virions can be budded from the infected cells, because of the genome-wide substitutions or the *crossover reading* between the two RNA strands by the reverse transcriptase in the host cells (Hu and Temin, 1990; Taylor *et al.*, 2008). These variant genomes have been classified into various types/groups and subtypes according to the general principle that sequences within any one subtype or sub-subtype should be more similar to each other than to sequences from the other subtypes (Foley *et al.*, 2018). Currently, HIV-1 is classified into M, N, and O groups, and the M group is further classified into subtypes A, B, C, D, F, G, H, J and K.

It has also been found that the two HIV RNA strands of a budded virion can be quite different when the host cell is infected by two different strains/subtypes of the virus (Taylor *et al.*, 2008). In fact, some budded virions have packed one RNA strand from each of the two different strains/subtypes of the virus (i.e., co-packed). When the co-packed genome enters a new cell, the crossover reading between this pair of RNAs by the reverse transcription can generate an interweaved genomic RNA composed of viral segments hopping cross the two co-packed subtypes (Hu and Temin, 1990). This newly packed viral genome (the two copies of the interweaved RNA) is termed as an inter-subtype recombinant genome. If an inter-subtype recombinant virus is transmitted to many people, it becomes one of the circulating strains in the HIV epidemic. The genome is then termed as a "circulating recombinant form (CRF)" to distinct from the pure subtypes.

Each CRF is labeled with an integer number and a combination of pure subtypes to indicate its chronological information and multiple genetic sources. These numbers follow a time order (chronological order) in which the CRFs first reached a detectable level in the population (Robertson *et al.*, 2000). For example, CRF05_DF means that this CRF is the fifth-earliest confirmed CRF epidemic having genetic sources from HIV-1 subtypes D and F, and there may be tens or hundreds of slightly mutated genome sequences labeled the same as CRF05_DF in a database. So far, there are 98 CRF types (different combinations of pure subtypes) documented at `https://www.hiv.lanl.gov/content/sequence/HIV/CRFs/CRFs.html`. These viral genomes not only recombine between viral RNAs from pure subtypes, but also recombine between those from pure subtypes and CRFs, exhibiting a high complexity of mutation and reproductive diversity. The virus is still spreading in the form of more complex CRFs (Zhang *et al.*, 2010). As precise subtyping has a crucial role in HIV therapy, knowing the correct subtypes of the patients can have a huge impact on the treatment outcome (Riemenschneider *et al.*, 2016b; Cashin *et al.*, 2015).

It is a challenging computational problem to make accurate identification for the genetic sources of a newly reported genome sequence of HIV-1. Current algorithms for the classification or prediction of HIV-1 subtypes and CRFs all focus on the prediction of one label for a given sequence. These methods include the adaptive context-based COMET (Struck *et al.*, 2014), the restriction fragment pattern-based CASTOR (Remita *et al.*, 2017), the phylogeny-based REGA (De Oliveira *et al.*, 2005), SCUEAL (Pond *et al.*, 2009), the k-mer-based KAMERIS (Solis-Reyes *et al.*, 2018), and the digital signal processing-based MLDSP (Randhawa *et al.*, 2019). These prediction methods are not capable of adequately identifying the multiple genetic sources of CRF sequences. Detailed comparative studies of these methods for the classification of infectious diseases, especially for the HIV viruses can be found in (Patiño-Galindo and González-Candelas, 2018; Fabeni *et al.*, 2017). More recently, a recombination analysis tool was developed to understand the status of viral recombination in the early stages of HIV infection (Song *et al.*, 2018). However, this tool does not provide automatic prediction or machine learning functions.

Our work exploits the complementary fit between *multi-label learning* and CRF *label set* prediction, and introduces a voting scheme to integrate various multi-label learning algorithms to strengthen the prediction of multiple genetic sources for a test sequence. The key idea of multi-label learning is to capture the unique sequence patterns from every pure subtype of genome sequences. When a test sequence contains multiple of these signature patterns, the algorithm can predict multiple labels for the test sequence. For example, the multi-label learning method Breknna (Spyromitros *et al.*, 2008) can predict multiple labels using iterative binary relevance and neighborhood distance to capture signature patterns. Another multi-label learning method MLKNN (Zhang and Zhou, 2007) adapts the traditional nearest neighbor approach to capture statistical information from the label sets of the neighboring sequences of a test sequence, and uses the maximum posteriori (MAP) principle to assign multiple labels. Multi-label ARAM (Benites and Sapozhnikova, 2015) uses neural networks for simultaneous prediction of multiple labels, which has been implemented in a Python environment (Szymański and Kajdanowicz, 2017). As these multi-label learning methods can capture the multiple signature patterns of CRF genomes from different angles, we use a voting of these prediction results to significantly firm the decision on the complete label set of a CRF genome. We note that multi-label learning and multi-task learning had been used for HIV-related research previously, for example to predict drug resistance (Gönen and Margoliny, 2014; Riemenschneider *et al.*, 2016a; Heider *et al.*, 2013), but not for the prediction of CRF multiple labels.

We extract two kinds of sequence features to represent every HIV-1 genome. Specifically, the genetic information of k-mers in the sequences are extracted using DLTree (Wu *et al.*, 2017) and the position information of k-mers in the sequences are extracted using PWKMER (Ma *et al.*, 2020). We select the most relevant features by assessing scores of the strings from a set of standard HIV-1 sequences (Thomas *et al.*, 2005) calculated by DLTree under various k settings. Important genetic features and the sequence position-related features (Ma *et al.*, 2020) are merged into a feature vector to describe the sequences for multi-label learning. In addition, we compute the pairwise distances of these sequences to form a distance matrix, and apply the classical Multi-Dimensional Scaling method to plot Molecular Distance Maps (Kari *et al.*, 2014) for better understanding the sequence clustering behavior of HIV-1 genomes.

Our prediction problem also involves the prediction of the time-order chronological number for a CRF sequence (e.g., the number "05" in CRF05_DF). The key idea is to calculate the chord distance between the test sequence and sequences in a reference set. The reference set of a pure subtype is defined as the whole set of the same pure subtype training data; while the reference set of a CRF is defined as those representative sequences from all the strains sharing the same CRF label set but having different chronological numbers. The final prediction is made based on a majority vote in the chord distance neighborhood of the test sequence.

2 Data sets

A genomic sequence data set of HIV-1 was downloaded from the Los Alamos (LANL) database (`https://www.hiv.lanl.gov/`) on May 12, 2019. The data set contains 7185 sequences, each labeled with the M group subtypes A, A1, A2, A3, A5, A6, A7, B, C, D, F1, F2, G, H, J, and K, or labeled with CRFs, where A1, A2, A3, A4, A5, A6 and A7 are the sub-subtypes of subtype A; F1 and F2 are the sub-subtypes of subtype F. The time-order chronological numbers of the CRFs are numbers from CRF01 to CRF98. The genetic source recombination in these CRF sequences occurred not only between the pure subtypes but also between the subtypes and partial CRFs. For example, CRF04_AGHKU is a very complicated CRF combining viral genetic sources from 5 subtypes of HIV-1. Such a complicated form of genomic type combination is sometimes shortened as CRF04_cpx.

The viral genome of HIV-1 consists of 9 genes: gag, pol, env, tat, rev, nef, vif, vpr, and vpu. In this work, we used the whole genome sequence or only the coding parts of the whole genome sequence (i.e., the coding sequences of all the 9 genes) to test the effectiveness of our method. As Hue *et al.* (2004) reported that the HIV-1 pol gene contains rich genetic information for distinguishing HIV-1 subtype sequences, we also considered using only the pol gene data for pure subtype and CRF label set prediction. The pol gene is located between positions 2000 and 5000 in the HIV-1 genome sequences, but the length and the starting position of the pol gene can be different in different genomes. In this work, we extracted

the coding sequences of the pol gene using the annotation information from all the genomes (i.e., from the '.gbk' file).

For prediction performance evaluation, the 7185 sequences were divided into a training set and a test set under a ratio 8:2. It is a random label-stratified division, i.e., every type of sequences follows the ratio 8:2. The training set contains 4431 pure-subtype sequences and 1362 CRF sequences, and the test set contains 1099 pure-subtype sequences and 293 CRFs. The pure subtype prediction and CRF prediction models were trained separately. The training set of the pure subtype prediction model only included 4431 pure subtype sequences, and the remaining 20% pure subtype sequences (i.e. 1099) were used as test set. Because most of the CRF sequence components are derived from the pure subtypes, the training set of the CRF prediction model contained both the 80% pure subtype sequences and 80% CRF sequences. After the CRF prediction model was constructed, the remaining 20% CRF sequences were used as test set.

All the sequences in the training set are of high sequence similarities. In fact, the maximal similarity of every training sequence within its own class of sequences is very close to its maximal similarity in the other classes of sequences. For example, there are 848 CRF sequences whose maximal similarity within their own class of sequences and whose maximal similarity in the other classes of sequences are close in a 3.0% similarity difference gap (such as 96.1% vs 94.1%, 94.7% vs 94.3%). Sometimes, the maximal similarity within their own class is lower than their maximal similarity in the other classes (such as 90.2% vs 90.4%, 95.7% vs 96.1%). This fact implies that there are only tiny difference between these subtypes of sequences. It is challenging for a learning model to capture the subtle unique patterns characterising each pure subtype or each CRF of HIV-1 sequences.

Each of the sequences in the test set can align to a highly similar sequence in its own training set and meanwhile can align to a highly similar sequence in the other classes of training data. The two maximum-similarity distributions of the CRF sequences (complete genome) from the test set are listed in Table 1. More details can be found at Supplementary Tables 2-4.

Table 1. Maximum-similarity distributions of 293 CRF complete genome sequences (from the test set which was obtained by random selection) with the sequences of their own type in the training set or with the sequences in the other types.

Max-similarity	<80%	80%-85%	85%-90%	90%-95%	>95%
with own type	3	0	2	66	222
in other types	4	0	64	223	2

In the performance evaluation, we removed 122 CRF complete genome sequences from the test set which are identical or nearly identical (99%) to a training sequence.

There is another way to remove the redundancy in this sequence set through the CD-HIT method (Li and Godzik, 2006). We set 95% as the similarity threshold for CD-HIT to cluster highly similar sequences. From the 3503 clusters, we random chose one to keep and the rest were removed. For these remaining sequences, we did a random selection of training sequences and test sequences, and then trained the multi-label learning model, and tested the model on the test data to get the accuracy performance.

To understand more about the prediction capacity of multi-label learning, apart from the random division of the data into a training set and a test set, we also purposely split the data according to the sequence-sampling year. In each CRF category, only those earlier-year sampled sequences (80%) were chosen as the training data, and the later-year sampled sequences (20%) were chosen as the test data set. The time information of the sampling years of all the sequences were obtained from `https://www.hiv.lanl.gov/`, and then they were sorted according to the sequence-sampling time. The two maximum-similarity distributions of these CRF sequences (complete genome) from the test set are listed in Table 2. More details can be found at Supplementary Tables 5, 6, and 7.

Table 2. Maximum-similarity distributions of 293 CRF complete genome sequences (from the test set which obtained by sampling year) with the sequences of their own type in the training set or with the sequences in the other types.

Max-similarity	<80%	80%-85%	85%-90%	90%-95%	>95%
with own type	0	1	9	146	137
in other types	1	3	61	228	0

Again, those test sequences having a maximum similarity higher than 99% with the training set were excluded from the prediction performance evaluation. For example, we removed 45 CRF complete sequences in this case.

3 Method

3.1 Overview

We combine three multi-label classification methods to predict the label set of a HIV-1 sequence. The method is able to predict one single label or predict multiple labels simultaneously for any test sequence. In the case when multiple labels are predicted, it means the virus is a CRF with genetic components from multiple subtypes.

The 7185 virus sequences are described using our newly constructed feature space. The feature space merges the frequency and position information of *k-mers* to capture sequence characteristics. The features underlining the frequency characteristics of *k-mers* are derived by DLTree (Wu *et al.*, 2017), and the feature information on the positions of *k-mers* are obtained by PWKMER (Ma *et al.*, 2020). In the process of acquiring relevant features, so-called standard *k-mers* are determined from 44 HIV-1 reference sequences to narrow down the feature space.

The work flow of our method is depicted at Fig 1. Details of the multi-label classification and feature space construction are presented in the following subsections.

3.2 Multi-label learning methods and voting for CRF label set prediction

Computational model construction of multi-label prediction is to obtain outstanding signature patterns from the training data of various labels. The prediction is to see how many signature patterns are contained in a test sequence. We take two approaches for multi-label prediction/classification, and vote their predictions for reliable decisions. The first approach for multi-label prediction is to convert the multi-label problem into multiple binary one-vs-other single-label prediction problem, e.g., Breknna (Zhang and Zhou, 2007; Benites and Sapozhnikova, 2015). The other is an algorithm-adaptive approach (Zhang and Zhou, 2007; Benites and Sapozhnikova, 2015) — it is about the adaptation of an existing single-label prediction method to deal with the simultaneous prediction of multiple labels, e.g., MLKNN (Zhang and Zhou, 2007; Benites and Sapozhnikova, 2015) and MLARAM (Benites and Sapozhnikova, 2015).

3.3 One-vs-other binary relevance for multi-label learning

Binary relevance (BR) is a binary classification of each label l ($l \in L$), i.e. the data $x \rightarrow \{l, \neg l\}$. BR converts the original data into $|L|$ number of data

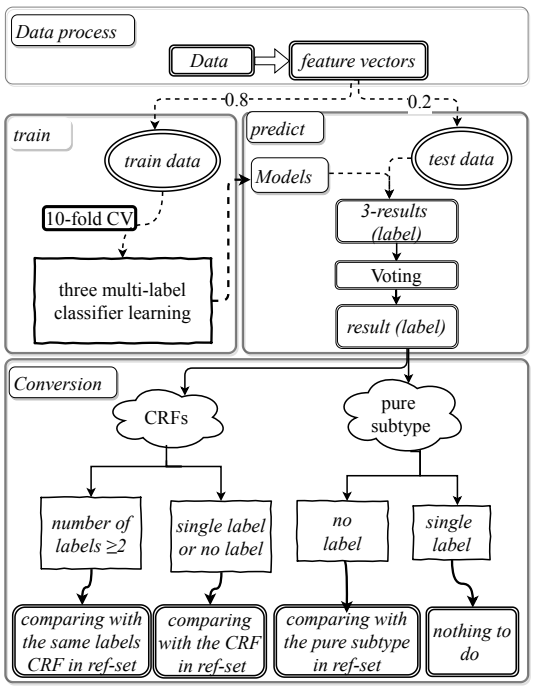

Fig. 1. Flowchart for analyzing HIV-1 sequences with multi-label learning, includes 'Data process', 'model training', 'prediction' and 'conversion'.

sets $D_l, l = 1, \ldots, |L|$. Breknna (Spyromitros *et al.*, 2008) combines the BR and k nearest neighbor to handle multi-label problems. The confidence c_l of every label l needs to be estimated in the process, by computing the percentage of every label among the k neighbors $N_x(k)$ of a given test instances x,

$$c_l = \frac{1}{k} \sum_{j=1}^{k} y_j(l) \tag{1}$$

where $y_j(l) = 1, j \in N_x(k)$, if j contains label l; otherwise $y_j(l) = 0$. If at least half of the neighbors contain labels $l \in L$, the highest confidence label is output as the label of x.

3.4 Adapted nearest neighbor classification for multi-label learning

The traditional k nearest neighbor (KNN) method can only predict a single label. MLKNN (Zhang and Zhou, 2007) adapts the KNN algorithm to handle multiple labels. During the training process, it is assumed that there are m data points in the training set and y is the label set. There are two kinds of events: (1) H_1^l: Contains the label l; (2) H_0^l: Does not contain the label l, where $l \in y$. Let $y_x(l)$ be the label vector of an instance $x \in m$, such that $y_x(l) = 1$, if x has the label l, and $y_x(l) = 0$ otherwise. Then,

- Let the prior probabilities of the two events in the m training data points be $P(H_1^l), P(H_0^l)$, where $P(H_0^l) + P(H_1^l) = 1$;
- Find the k-nearest neighbours for a test instance t in the training data and calculate the posterior probability of the two events. First, finding the k nearest neighbors of t, and calculating the number of the nearest neighbor of t containing label l, $C_t(l)$,

$$C_t(l) = \sum_{a \in N(t)} y_t(l). \tag{2}$$

  Then, estimating the posterior probability of the k neighbors of t for the two events based on the statistical inclusion of the number of labels l: $P(E_j^l|H_1^l)$ and $P(E_j^l|H_0^l)$, where $E_j^l (j \in \{1, \cdots, k\})$ means that there are j points having the label l in the k nearest neighbors;
- Use maximum posteriori principle to identify whether the test data t contains label l.

$$y_t(l) = arg\ max_{b \in \{0,1\}} P(H_b^l | E^l_{C_t(l)}) \tag{3}$$

Using the Bayesian rules, $y_t(l)$ can be derived as:

$$\begin{aligned} y_t(l) &= arg\ max_{b \in \{0,1\}} \frac{P(H_b^l) P(E^l_{C_t(l)} | H_b^l)}{P(E^l_{C_t(l)})} \\ &= arg\ max_{b \in \{0,1\}} P(H_b^l) P(E^l_{C_t(l)} | H_b^l) \end{aligned} \tag{4}$$

MLARAM is another multi-label learning methods which adapts the idea of neural learning. We refer to (Benites and Sapozhnikova, 2015) for MLARAM's complicated learning process and details.

3.5 Voting of predictions from various multi-label learning methods

Given a test sequence $newS$, by the prediction of the above three multi-label learning algorithms, three sets of predicted pure-subtype labels can be obtained. Some of the sets may contain only one label, the other may contain multiple labels, and some of the three sets may be the same. We use majority voting to make a collective decision on the final labels of the test sequence $newS$. The voting can effectively avoid possible poor prediction performance by a single method.

Let $label(newS, \text{Breknna})$ be the set of labels predicted by Breknna; $label(newS, \text{MLKNN})$ be the set of labels predicted by MLKNN; and $label(newS, \text{MLARAM})$ be the set of labels predicted by MLARAM. We define $label(newS, \text{voting})$ as the set of labels which belongs to at least two of the three label sets: $label(newS, \text{Breknna})$, $label(newS, \text{MLKNN})$, and $label(newS, \text{MLARAM})$. That is, only when labels predicted by at least two models are exactly the same, they are recommended as the final labels of the test sequence $newS$.

3.6 Prediction of chronological number

Denote $label(newS, \text{voting})$ as crf which is a subset of the pure subtypes of HIV-1 genome sequences. Suppose in the training data, crf-labeled sequences have w chronological numbers $n_1, \ldots, n_w$. From each of these w groups of training sequences, randomly choose v number of representative sequences, and combine them into a reference sequence set REF. Calculate the chord distance of $newS$ with every sequence in REF, and identify those sequences in REF that have the closest distance with $newS$. After this distance sorting, we use voting on the chronological numbers in the tail part of the list of sequences to decide the chronological number for $newS$.

3.7 Feature space construction to describe HIV-1 sequences

3.7.1 Features based on genetic information of *k*-mers

A *k-mer*-based alignment-free method was first proposed by (Blaisdell, 1986) to compute the similarity among sequences. A dynamical language based approach (called *DLTree*), which is also a *k-mer*-based method, was applied in phylogeny reconstruction by (Wu *et al.*, 2017). In the process of molecular evolution, the genetic information contained in molecular sequence is often masked by noise (Charlebois and Beiko, 2003). Therefore, we need to remove the random background for accurately identifying the phylogenetic relationship between sequences. DLTree simulates the random background of sequences using a dynamical language model. The genetic information of s which is extracted

by DLTree, is referred as $q_k(s)$ (more details are provided in the supplementary material). Hence we can convert a sequence into a vector 1×4^k with the components $q_k(s)$, as derived by the DLTree method.

The position information of nucleotide bases is significant in genetics as shown by (Ding *et al.*, 2013; Ma *et al.*, 2020), where phylogenetic analysis were conducted using the information from the positions of k-mers. These features are also used here together with the k-mer frequency features to describe HIV-1 sequences for subtype prediction. We denote the position information of s as $F(s)$, and provide the details in the supplementary material. In summary, we use two different perspectives to extract k-mer genetic information (i.e., F and q_k), and combine them to form our feature space. However, if all k-mers are used directly, our feature dimension is very large, so we attempted to reduce the dimension of the feature space without losing much information.

3.7.2 Using standard *k-mer*s to narrow down the feature space

To reduce the dimension of this vector, we use a HIV-1 reference sequence set to narrow down the search space of k-mers and choose good ones as outstanding features (Wu *et al.*, 2007). The LANL sequence database provides 44 reference sequences (Thomas *et al.*, 2005) for our dimension reduction. The detailed scoring criteria for selecting outstanding k-mers are as follows:

- Combine the 44 sequences into one super-genome.
- Use relative entropy to score the *k-mers*:

$$Z(\alpha) = \sum_{i=1}^{44} |S(\alpha, i)| \log_2 |\frac{S(\alpha, i)}{S(\alpha)}| \tag{5}$$

α is a *k-mer* $(3 \leq k \leq kmax)$, $Z(\alpha)$ is the score of α, $S(\alpha, i)$ is $q_k(\alpha)$ in the i-th reference sequence $(i = 1, \cdots, 44)$ and $S(\alpha)$ is $q_k(\alpha)$ in the super-genome obtained by the DLTree method.

When $k = 10$, there are $4^{10} = 1048576$ k-mers. When k steps from 1 to 10, there are total $4^1 + 4^2 + \cdots + 4^{10} = 1398100$ k-mer strings. We sort all of these strings according to their Z scores into a descending order, and select the first 5000 strings (the length k of which may be different) as standard k-mers. We extract the evolution information for these 5000 standard k-mers only in order to reduce the dimension of the feature vector.

3.8 Visualization of clusters of HIV-1 genome sequences by 3D MoDMaps

By our feature extraction methods, a sequence is converted into a vector with 10,000 dimensions (i.e., 5000 $q_k(s)$ and 5000 $F(s)$ with regard to the 5000 standard k-mers). We calculate pairwise cosine distances of these vectors to form a distance matrix. The Molecular Distance Maps (MoDMaps) (Kari *et al.*, 2014; Randhawa *et al.*, 2019) is applied to this distance matrix to visualize these sequences in a 3D space, where every dot represents a sequence. The distance between points in the 3D space is almost identical to the elements in the corresponding distance matrix (Kari *et al.*, 2014) under the classical Multi-Dimensional Scaling. MoDMaps are effective for understanding clustering behavior of the genome sequences.

4 Results

We show the high quality of the top-ranked features for the distinction of HIV-1 genome sequences through the construction of phylogenetic trees and 3D MoDMaps. We present detailed performance on CRF label predictions achieved by the multi-label learning algorithms, especially results from the majority voting. We also report the performance of two state-of-the-art single-label prediction methods to illustrate the superiority of multi-label learning to cope with the CRF label set prediction problem.

4.1 High quality of the top-ranked features for the construction of phylogenetic trees

Distribution of k-mer features. According to the above Z scores of k-mers $(3 \leq k \leq 10)$, we sorted all these strings into a decreasing order. We selected the first 500 and 5000 strings (their lengths may be quite different). The proportions of the strings with different lengths in the first 500 or in the first 5000 strings are showed in Table 3. The percentage of the k-mers with $k = 8$ in the first 500 or in the first 5000 strings is the highest, followed by length-9 and length-7 k-mers. Length-5 and length-6 k-mers can be also top-ranked, although their proportion is small. This statistics suggests that top-ranked features can come from different k-mer slots. Single-length k-mers as features may be inadequate for good classification.

Table 3. The distribution of top-ranked strings with k up to 10

k	5	6	7	8	9	10
Top 500	0.2%	0.2%	19.4%	**41%**	29%	10.2%
Top 5000	0.08%	1.8%	15.96%	**38.54%**	30.96%	12.66%

Clear phylogenetic tree of HIV-1 sequences. To test the quality of the genetic information possessed by these features, the 44 pure-subtype reference sequences, 4 CRF01 sequences, and 4 CRF02 sequences were used as input to construct a phylogenenic tree. The $q_k(s)$ vector of these 52 sequences was extracted with the 5000 standard k-mers. The distance between two sequences was calculated by the definition of chord distance, and the phylogenetic tree was built by Mega7 (Kumar *et al.*, 2016). From the tree structure (Fig 2), the different M, N and O groups are all clearly and correctly distinguished. The subtypes B and D close in biology are depicted in neighbourhood in the tree as well. The sequences belonging to CRF01 or to CRF02 are clustered together respectively in the tree and are close to subtype A. Overall, sequences belonging to different subtypes in the M group are clearly separated in the tree. These results demonstrate that our top-ranked features are of high quality; they separate different types of sequences and keep the right phylogeny of the sequences.

4.2 Distinction of CRF sequences by 3D MoDMaps

We performed cluster analysis of the 10000-dimension feature vectors in the test set using MoDMaps (Kari *et al.*, 2014). Fig 3 (1) and (2) plot the 3D MoDMaps of the complete genomes and the pol genes of the pure subtype sequences in the test set, respectively. Fig 4 (1) and (2) depict the 3D MoDMaps of the complete genomes and the pol genes for the CRF sequences in the test set, respectively. Due to the big variety of CRF types, we only selected those types with a relatively large number of sequences to plot 3D MoDMaps. As seen in Fig 3 (1) and (2), subtype B and subtype D are closer together than with other subtypes, consistent with the real situation. In Fig 4 (1) and (2), '07' and '08' have the closest distance, because they have the same labels (both 'B' and 'C'). Overall, the sequences of the same type are well clustered in 3D MoDMaps using the 10000-dimension feature vectors which we extracted. In addition, close relationships between some CRFs and their pure subtypes can be observed from the 3D MoDMaps (see details at the supplementary material).

4.3 Almost perfect label-set prediction by multi-label learning

The first three rows of Table 4 are the prediction accuracies on the test set (randomly selected sequences) achieved by Breknna, MLKNN, and MLARAM. The 'voting' line is the accuracy achieved by our voting approach. Symbols 'cg_pure' and 'cg_crf' stand for the complete genome pure subtype sequences and complete genome CRF sequences,

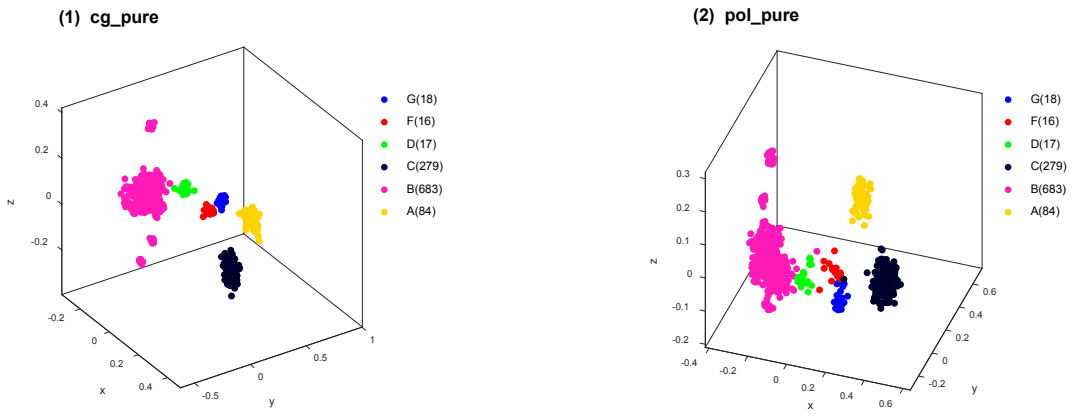

Fig. 3. 3D MoDMaps visualization of the complete genomes and the pol genes of the HIV-1 pure subtypes. (1) is the complete genome pure sequence (cg_pure), and (2) is the pol gene pure sequence (pol_pure).

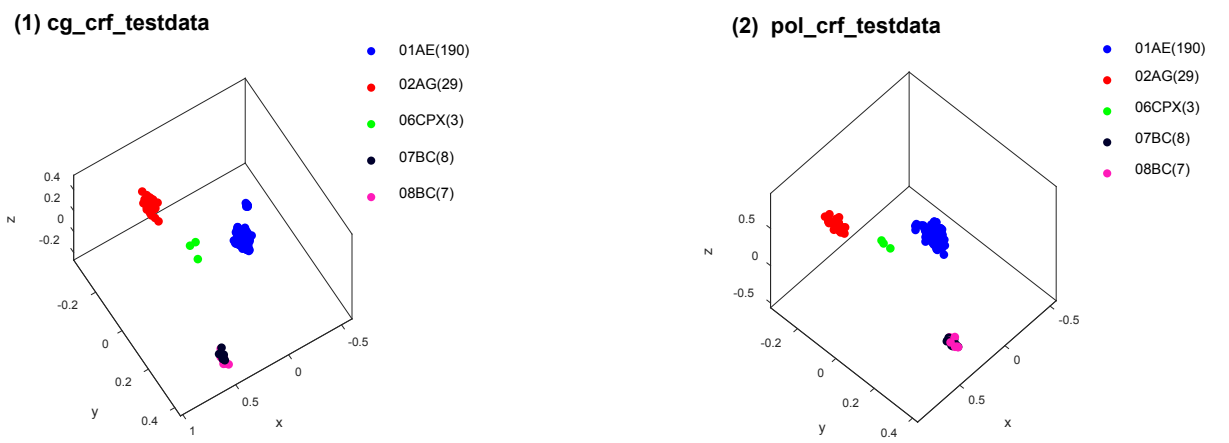

Fig. 4. 3D MoDMaps visualization of the complete genomes and pol genes of the HIV-1 CRF sequences. (1) is the complete genome CRF sequence (cg_crf), and (2) is the pol gene CRF sequence (pol_crf).

respectively. Similarly, 'pol_pure' and 'pol_CRF' represent that the pure subtype sequences and CRF sequences are limited to the pol gene coding sequence, respectively. *'Label'* means that the prediction is focused on the pure label or multi-label prediction (e.g., simultaneous prediction of A, D and F), while *'Type'* means that the prediction is focused on label-set prediction and chronological number prediction together (e.g., simultaneous prediction of A, D and F, and prediction of the chronological number 05).

These multi-label learning methods have made almost perfect prediction accuracies either when the feature space is generated from the complete genome sequences or generated from the pol gene coding sequences. Although the accuracies provided by some classifiers are about 91.81%, by voting the predicted labels from the three models, the prediction accuracy is stable and is always above 95.32%. These results confirm that multi-label learning is the right choice of machine learning approaches for CRF label-set prediction, signifying the contribution of this research work. Our voting scheme combines the advantages of the three methods which can effectively avoid the biased preferences made by an individual model, for example, the Breknna's prediction on cg_CRF.

Table 4. Prediction accuracies by multi-label learning algorithms Breknna, MLKNN, and MLARAM as well as their voting performance for HIV-1 sequences.

	classifier	cg_pure	cg_CRF	pol_pure	pol_CRF
Label	Breknna	0.9982	0.9181	0.9805	0.9667
	MLKNN	0.9963	0.9766	0.9805	0.9867
	MLARAM	0.9963	0.9474	0.9941	0.96
	Voting	0.9963	0.9532	0.9824	0.98
Type	Breknna	0.9982	0.9591	0.9883	0.9733
	MLKNN	0.9963	0.9708	0.9922	0.98
	MLARAM	0.9963	0.9708	0.9961	0.9667
	Voting	0.9963	0.9708	0.9902	0.9733

For the redundancy-removed data set by CD-HIT, the number of complete genomes is 3503 (CRF:866, pure subtype: 2637) and the number of pol-gene sequences is 1474 (CRF:259, pure subtype :1215). We randomly selected sequences as training and test under each type

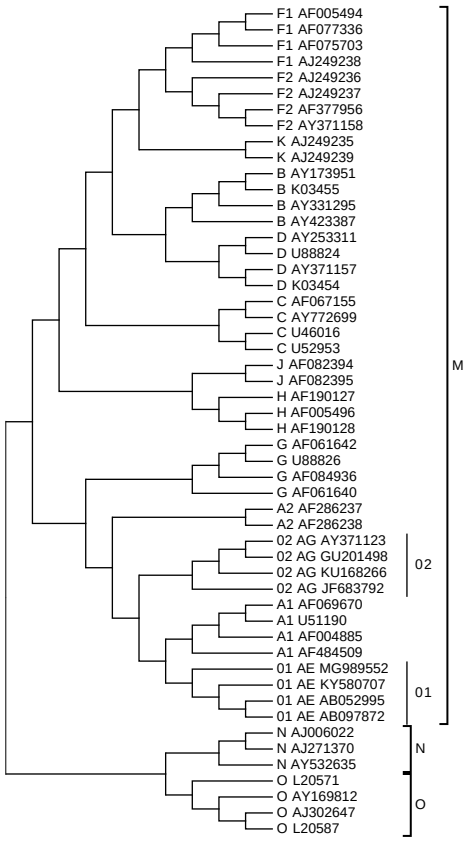

Fig. 2. A phylogenetic tree of the 44 reference sequences, 4 sequences from CRF01 and 4 sequences from CRF02 using the top-ranked 5000 reference k-mer strings.

according to the ratio of 8:2. Table 5 shows the test performance which is also excellent.

Table 5. Prediction accuracies by multi-label learning algorithms Breknna, MLKNN, and MLARAM as well as their voting performance for the CD-HIT clustered HIV-1 sequences.

	classifier	cg_pure	cg_CRF	pol_pure	pol_CRF
Label	Breknna	0.9981	0.8963	0.8950	0.7826
	MLKNN	0.9962	0.9778	0.9664	0.7826
	MLARAM	0.9962	0.9852	0.9748	0.8696
	Voting	0.9962	0.9778	0.9622	0.7826
Type	Breknna	0.9981	0.9481	0.9034	0.9130
	MLKNN	0.9981	0.9704	0.9664	08696
	MLARAM	0.9962	0.9778	0.9790	0.8696
	Voting	0.9981	0.9704	0.9748	0.9130

On the test set sorted according to the time information of the sequence-sampling year, the prediction performance is even better than on the randomly selected test sequences. See Table 6 for the almost perfect prediction accuracies by the multi-label learning algorithms and their voting scheme.

Table 6. Prediction accuracies by multi-label learning algorithms Breknna, MLKNN, and MLARAM as well as their voting performance (on the test set which was constructed according to the sequence-sampling year). The bold font indicates the best prediction performance by this work.

	classifier	cg_pure	cg_CRF	pol_pure	pol_CRF
Label	BREKNNA	0.9990	0.9234	0.9827	0.9785
	MLKNN	0.9990	0.9677	0.9798	0.9828
	MLARAM	0.9962	0.9879	0.9990	0.9828
	Voting	0.9990	0.9718	0.9866	0.9828
Type	BREKNNA	0.9990	0.9677	0.9866	0.9914
	MLKNN	1.0000	0.9839	0.9904	0.9914
	MLARAM	0.9962	0.9919	0.9990	0.9914
	Voting	**1.0000**	**0.9839**	**0.9914**	**0.9914**

4.4 Indirect comparison with state-of-the-art single-label prediction performance

Two current methods COMET (Struck *et al.*, 2014) and MLDSP (Randhawa *et al.*, 2019) are both single-label prediction methods for HIV-1 subtype sequences (i.e., one label predicted for each test sequence). They label the CRF sequences all with 'crf'. But, we predict detailed genetic sources for a CRF sequence. COMET is a web based tool which only accepts uploading of test data, 800 sequences each time. The user is not allowed to re-train the model. MLDSP is a stand-alone tool. We used the training data of the current work to re-train the model, and evaluate the performance on the test data.

Tables 7, 8, and 9 show the prediction accuracy of COMET and MLDSP on the three test sets. Under an indirect comparison (COMET and MLDSP predict the single label 'crf' versus we predict the complete label-set of each CRF sequence), our method is much superior to COMET and MLDSP no matter the feature space is generated from whole genome sequence or from the pol gene coding sequence for the CRF label set prediction. As an additional analysis, we found that MLDSP's performance could be slightly improved when our feature vectors were used for the pol gene sequences. See details at the Supplementary material.

Table 7. Indirect comparison with two single-label prediction methods based on the randomly divided data.

method	cg_pure	cg_CRF	pol_pure	pol_CRF
COMET	0.9191	0.8070	0.9531	0.92
MLDSP	0.9596	0.7427	0.9492	0.7467
Our method	**0.9963**	**0.9708**	**0.9902**	**0.9733**

Table 8. Indirect comparison with two single-label prediction methods on the test set after removing the redundancy by CD-HIT

method	cg_pure	cg_CRF	pol_pure	pol_CRF
COMET	0.9328	0.8444	**0.9832**	0.7826
MLDSP	0.9501	0.7704	0.9160	0.6957
Our method	**0.9981**	**0.9704**	0.9748	**0.9130**

Table 9. Indirect comparison with two single-label prediction methods on the test set sorted by the time information of sequence-sampling years.

method	cg_pure	cg_CRF	pol_pure	pol_CRF
COMET	0.9428	0.7984	0.9702	0.9099
MLDSP	0.9552	0.7863	0.9597	0.8197
Our method	**1.0000**	**0.9839**	**0.9914**	**0.9914**

5 Discussions on our wrong predictions

A small number of wrong predictions have been made by our multi-label learning method for the labels of pure-type sequences or recombinant sequences. This section provides details about which sequences are prone of wrong prediction, whether the domain database is a gold standard, and why CRF prediction is really a challenging problem.

Where we have made wrong predictions? In the label prediction for pure subtype sequences, we found that most of the wrong predictions are that subtype A was predicted as a sub-subtype of A, or the opposite. For example, 'A1_JQ403028' was predicted as 'A', and 'A_DQ396400' was predicted as 'A1' . A very small number of subtype D sequences were predicted as subtype B, and some subtype B sequences wrongly predicted as D. In the label set prediction of CRF sequences, a majority of the wrong predictions is that the set of predicted labels is incomplete. Some of the predicted label sets contains only one label, but the complete set of labels of the sequence actually contains more than that. For example, the labels of 'CRF15_01B' are '01' and 'B', but our prediction result is '01' (or a combination of '01' and a wrong label). Other wrong predictions include the whole set of labels of a CRF sequence being correctly predicted, but its chronological number was wrongly predicted. For example, in the actual prediction, some '70_BF1' sequences are predicted to be '71_BF1'.

Why HIV-1 CRF prediction is really difficult? When both of training and test data sets contain only pure subtype sequences, the label prediction problem is easy, and the prediction performance is excellent. But, simultaneous and complete prediction of the multiple genetic sources of a CRF sequence is really challenging, especially when the chronological number is also required to be assigned. To understand this point, we searched the accession numbers of those sequences which were wrongly predicted by all of COMET, MLDSP and our multi-label voting method. Fortunately, there are only two sequences in this list. They are CRF11_cpx (with the accession number KP718935) and CRF15_01B (with the accession number DQ354120). In other words, the predicted sets of labels for these two sequences by all of the three methods are inconsistent with the records in the HIV database. In fact, CRF11_cpx_KP718935 (Montavon *et al.*, 2002) has a complicated label set containing subtypes A, G, J and CRF01_AE. See Fig 5 for a schematic representation of the mosaic structure of the CRF11_cpx genome, where 5 different subtypes of small rectangles are indeed intricately composed to build CRF11_cpx.

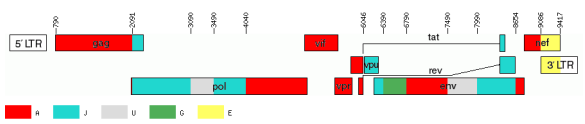

Fig. 5. Schematic representation of the mosaic structure of the CRF11_cpx genome, download from LANL(https://www.hiv.lanl.gov/)

COMET could not predict its type; the predicted result by MLDSP was 'CRF01'; and the label set predicted by our multi-label voting method contains '01' and 'B' only. It is really sometimes difficult to predict the complicated composition of a CRF sequence by machine learning methods as indicated by this example.

Whether an expert-maintained domain database is a gold benchmark? For the pol gene of sequence DQ354120 (CRF15_01B), all of the three methods (COMET, MLDSP, and our voting) predicted it as 'CRF_01'. Although this prediction result is not consistent with the database record (i.e. CRF15_01B), we downloaded the schematic representation of the mosaic structure (Fig 6) of CRF15_01B from LANL (https://www.hiv.lanl.gov/) to understand whether the database has an error. As the mosaic structure clearly shows that only CRF_01 is on the pol gene, we believe that all the predictions are consistent with the mosaic structure. This implies that the database need to update the label information for the pol gene of sequence to avoid confusion with the whole sequence labels.

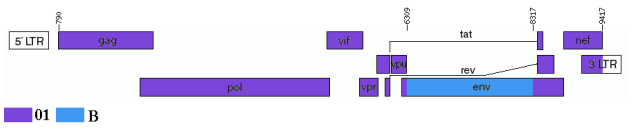

Fig. 6. Schematic representation of the mosaic structure of the CRF15_01B genome, downloaded from LANL (https://www.hiv.lanl.gov/)

6 Conclusion

Over time, HIV-1 genomes not only combine between viral RNAs from pure subtypes, but also recombine between those from pure subtypes and CRFs, exhibiting a high complexity of mutation and reproductive diversity. We propose using multi-label learning to capture the patterns of pure subtypes and use voting of various multi-label learning algorithms to strengthen the prediction of complete multiple genetic sources of a CRF. In the step of constructing the feature space, we extract both genetic and position information of k-mers as a merged feature vector to describe every HIV-1 genome sequence. This prediction problem also involves the prediction of the time-order chronological number for a CRF sequence. We solved the problem using a majority vote in the chord distance neighborhood of the test sequence. Our extensive experiments have demonstrated that the top-ranked features are very effective for building clear phylogenetic trees of HIV-1 genomes and also excellent for depicting clear 3D MoDMaps to understand the sequence clustering behavior of HIV-1 genomes. The multi-label learning algorithms with the top-ranked features have provided almost-perfect accuracies for the complete prediction of multiple genetic sources of CRF sequences. The performance is much superior to the best existing methods under an indirect comparison. We have also conducted an analysis on which of the sequences prone of wrong prediction are, and why the CRF label set prediction is indeed challenging, and a more interesting question on how to identify mislabeled record in the domain-expert maintained database through machine learning. As a future work, we will conduct deep analysis on the unique patterns and sequence motifs corresponding to each pure subtypes of HIV-1 genomes.

Acknowledgements

We thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of our manuscript. We greatly thank to Associate Professor Guosheng Han and Dr. Qi Wu for helpful discussions.

Funding

This work has been supported by National Natural Science Foundation of China (Grant No. 11871061); Collaborative Research project for Overseas Scholars (including Hong Kong and Macau) of National Natural Science Foundation of China (Grant No. 61828203).

Conflict of Interest

None declared.

References

Benites, F. and Sapozhnikova, E. (2015). Haram: A hierarchical aram neural network for large-scale text classification. In *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*, pages 847–854.

Blaisdell, B. E. (1986). A measure of the similarity of sets of sequences not requiring sequence alignment. *Proceedings of the National Academy of Sciences*, **83**(14), 5155–5159.

Cashin, K., Gray, L. R., Harvey, K. L., Perez-Bercoff, D., Lee, G. Q., Sterjovski, J., Roche, M., Demarest, J. F., Drummond, F., Harrigan, P. R., Churchill, M. J., and Gorry, P. R. (2015). Reliable genotypic tropism tests for the major HIV-1 subtypes. *Scientific Reports*, **5**, 21–23.

Charlebois, R. L. and Beiko, Robert G.and Ragan, M. A. (2003). Microbial phylogenomics: Branching out. *Nature*, **421**.

De Oliveira, T., Deforche, K., Cassol, S., Salminen, M., Paraskevis, D., Seebregts, C., Snoeck, J., Van Rensburg, E. J., Wensing, A. M., Van De Vijver, D. A., *et al.* (2005). An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**(19), 3797–3800.

Ding, S., Li, Y., Yang, X., and Wang, T. (2013). A simple k-word interval method for phylogenetic analysis of DNA sequences. *Journal of Theoretical Biology*, **317**, 192–199.

Fabeni, L., Berno, G., Fokam, J., Bertoli, A., Alteri, C., Gori, C., Forbici, F., Takou, D., Vergori, A., Zaccarelli, M., Maffongelli, G., Borghi, V., Latini, A., Pennica, A., Mastroianni, C. M., Montella, F., Mussini, C., Andreoni, M., Antinori, A., Perno, C. F., and Santoro, M. M. (2017). Comparative Evaluation of Subtyping Tools for Surveillance of Newly Emerging HIV-1 Strains. *Journal of Clinical Microbiology*, **55**, 2827–2837.

Foley, B. T., Korber, B. T. M., Leitner, T. K., Apetrei, C., Hahn, B., Mizrachi, I., Mullins, J., Rambaut, A., and Wolinsky, S. (2018). Hiv sequence compendium 2018. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Gönen, M. and Margoliny, A. A. (2014). Drug susceptibility prediction against a panel of drugs using kernelized Bayesian multitask learning. *Bioinformatics*, **30**(17), 556–563.

Heider, D., Senge, R., Cheng, W., and Hüllermeier, E. (2013). Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics*, **29**(16), 1946–1952.

Hu, W. S. and Temin, H. M. (1990). Genetic consequences of packaging two RNA genomes in one retroviral particle: Pseudodiploidy and high rate of genetic recombination. *Proceedings of the National Academy of Sciences of the United States of America*, **87**(4), 1556–1560.

Hue, S., Clewley, J., Cane, P., and Pillay, D. (2004). Hiv-1 pol gene variation is sufficient for reconstruction of transmissions in the era of antiretroviral therapy. *AIDS (London, England)*, **18**, 719–28.

Kari, L., Hill, K. A., Sayem, A. S., Karamichalis, R., Bryans, N., Davis, K., and Dattani, N. S. (2014). Mapping the space of genomic signatures. *Plos One*, **10**(5), e0119815.

Kumar, S., Stecher, G., and Tamura, K. (2016). Mega7: Molecular evolutionary genetics analysis version 7.0 for bigger datasets. *Molecular Biology & Evolution*, **33**(7), 1870.

Li, W. and Godzik, A. (2006). Cd-hit: A fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, **22**(13), 1658–1659.

Ma, Y., Yu, Z., Tang, R., Xie, X., Han, G., and Anh, V. V. (2020). Phylogenetic analysis of hiv-1 genomes based on the position-weighted k-mers method. *Entropy*, **22**(2), 255.

Montavon, C., Vergne, L., Bourgeois, A., Mpoudi-Ngole, E., Malonga-Mouellet, G., Butel, C., Toure-Kane, C., Delaporte, E., and Peeters, M. (2002). Identification of a new circulating recombinant form of hiv type 1, CRF11-cpx, involving subtypes A, G, J, and CRF01-AE, in central africa. *AIDS Research and Human Retroviruses*, **18**(3), 231–236.

Moutouh, L., Corbeil, J., and Richman, D. D. (1996). Recombination leads to the rapid emergence of hiv-1 dually resistant mutants under selective drug pressure. *Proceedings of the National Academy of Sciences*, **93**(12), 6106–6111.

Patiño-Galindo, J. and González-Candelas, F. (2018). Molecular evolution methods to study HIV-1 epidemics. *Future Virology*, **13**(6), 399–404.

Pond, S. L., Posada, D., Stawiski, E., Chappey, C., Poon, A. F., Hughes, G., Fearnhill, E., Gravenor, M. B., Brown, A. J., and Frost, S. D. (2009). An evolutionary model-based algorithm for accurate phylogenetic breakpoint mapping and subtype prediction in HIV-1. *PLoS Computational Biology*, **5**(11).

Rajarapu, G. (2014). Genes and Genome of HIV-1. *Journal of Phylogenetics & Evolutionary Biology*, **02**(01), 1–7.

Rambaut, A., Posada, D., Crandall, K. A., and Holmes, E. C. (2004). The causes and consequences of HIV evolution. *Nature Reviews Genetics*, **5**(1), 52–61.

Randhawa, G. S., Hill, K. A., and Kari, L. (2019). ML-DSP: Machine Learning with Digital Signal Processing for ultrafast, accurate, and scalable genome classification at all taxonomic levels. *BMC Genomics*, **20**(1), 1–21.

Remita, M. A., Halioui, A., Malick Diouara, A. A., Daigle, B., Kiani, G., and Diallo, A. B. (2017). A machine learning approach for viral genome classification. *BMC Bioinformatics*, **18**(1), 1–11.

Riemenschneider, M., Senge, R., Neumann, U., Hüllermeier, E., and Heider, D. (2016a). Exploiting HIV-1 protease and reverse transcriptase cross-resistance information for improved drug resistance prediction by means of multi-label classification. *BioData Mining*, **9**(1), 1–6.

Riemenschneider, M., Cashin, K. Y., Budeus, B., Sierra, S., Shirvani-Dastgerdi, E., Bayanolhagh, S., Kaiser, R., Gorry, P. R., and Heider, D. (2016b). Genotypic Prediction of Co-receptor Tropism of HIV-1 Subtypes A and C. *Scientific Reports*, **6**, 1–9.

Robertson, D. L., Anderson, J., Bradac, J., Carr, J., Foley, B., Funkhouser, R., Gao, F., Hahn, B., Kalish, M., Kuiken, C., *et al.* (2000). Hiv-1 nomenclature proposal. *Science*, **288**(5463), 55–55.

Solis-Reyes, S., Avino, M., Poon, A., and Kari, L. (2018). An open-source k-mer based machine learning tool for fast and accurate subtyping of HIV-1 genomes. *PloS one*, **13**(11), e0206409.

Song, H., Giorgi, E. E., Ganusov, V. V., Cai, F., Athreya, G., Yoon, H., Carja, O., Hora, B., Hraber, P., Romero-Severson, E., Jiang, C., Li, X., Wang, S., Li, H., Salazar-Gonzalez, J. F., Salazar, M. G., Goonetilleke, N., Keele, B. F., Montefiori, D. C., Cohen, M. S., Shaw, G. M., Hahn, B. H., McMichael, A. J., Haynes, B. F., Korber, B., Bhattacharya, T., and Gao, F. (2018). Tracking HIV-1 recombination to resolve its contribution to HIV-1 evolution in natural infection. *Nature Communications*, **9**(1).

Spyromitros, E., Tsoumakas, G., and Vlahavas, I. (2008). An empirical study of lazy multilabel classification algorithms. *Proc. 5th Hellenic Conference on Artificial Intelligence (SETN 2008)*.

Struck, D., Lawyer, G., Ternes, A. M., Schmit, J. C., and Bercoff, D. P. (2014). COMET: Adaptive context-based modeling for ultrafast HIV-1 subtype identification. *Nucleic Acids Research*, **42**(18), 1–11.

Szymański, P. and Kajdanowicz, T. (2017). A scikit-based Python environment for performing multi-label classification. *ArXiv e-prints*.

Taylor, B. S., Sobieszczyk, M. E., Mccutchan, F. E., and Hammer, S. M. (2008). The Challenge of HIV-1 Subtype Diversity Origin of HIV a nd Mechanisms of HIV Diversity. *The New England Journal of Medicine*, **15**(10), 1.

Thomas, L., Bette, K., Marcus, D., Charles, C., and Brian, F. (2005). Hiv-1 subtype and circulatinerinant form (crf) reference sequences. accessible through http://www.hiv.lanl.gov/content/ hiv-db/reviews/refseqs2005/refseqs05.html.

Wu, Q., Yu, Z. G., and Yang, J. (2017). Dltree: efficient and accurate phylogeny reconstruction using the dynamical language method. *Bioinformatics*, **33**(14).

Wu, X., Cai, Z., Wan, X. F., Hoang, T., Goebel, R., and Lin, G. (2007). Nucleotide composition string selection in HIV-1 subtyping using whole genomes. *Bioinformatics*, **23**(14), 1744–1752.

Zhang, M., Foley, B., Schultz, A. K., Macke, J. P., Bulla, I., Stanke, M., Morgenstern, B., Korber, B., and Leitner, T. (2010). The role of recombination in the emergence of a complex and dynamic HIV epidemic. *Retrovirology*, **7**(May 2014).

Zhang, M.-L. and Zhou, Z.-H. (2007). Ml-knn: A lazy learning approach to multi-label learning. *Pattern recognition*, **40**(7), 2038–2048.