WHOLE GENOME COMPOSITION DISTANCE FOR HIV-1 GENOTYPING

Xiaomeng Wu, Randy Goebel
Department of Computing Science, University of Alberta
Edmonton, Alberta T6G 2E8, Canada
Email: xiaomeng, geobel@cs.ualberta.ca

Xiu-Feng Wan
Systems Biology Laboratory, Department of Microbiology, Miami University
Oxford, OH 45056, USA.
E-mail: wanx@muohio.edu

Guohui Lin*
Department of Computing Science, University of Alberta
Edmonton, Alberta T6G 2E8, Canada
Email: ghlin@cs.ualberta.ca

Existing HIV-1 genotyping systems require a computationally expensive phase of multiple sequence alignments and the alignments must have a sufficiently high quality for accurate genotyping. This is particularly a challenge when the number of strains is large. Here we propose a whole genome composition distance (WGCD) to measure the evolutionary closeness between two HIV-1 whole genomic RNA sequences, and use that measure to develop an HIV-1 genotyping system. Such a WGCD-based genotyping system avoids multiple sequence alignments and does not require any pre-knowledge about the evolutionary rates. Experimental results showed that the system is able to correctly identify the known subtypes, sub-subtypes, and individual circulating recombinant forms.

Keywords: String composition; Whole genome phylogenetic analysis; Neighbor joining; HIV-1 genotyping; Circulating recombinant form

1. INTRODUCTION

Acquired Immune Deficiency Syndrome (AIDS) is caused by a virus known as human immunodeficiency virus (HIV). The first case of AIDS was reported in the United States in 1981 and has since become a major worldwide epidemic. By the end of 2005, over 900,000 people had been diagnosed with HIV infection, and more than 2.3 million were estimated to be HIV positive (`http://www.who.it`).

There are two types of HIV: HIV-1 and HIV-2. HIV-1 is more pathogenic than HIV-2, and most of HIV infection is caused by HIV-1 [1]. HIV is a retrovirus, which has one RNA genome segment encoding 9 genes, including *env*, *gag*, *nef*, *pol*, *rev*, *tat*, *vif*, *vpr*, and *vpu* (for HIV-1) or *vpx* (for HIV-2). Similar to other RNA viruses, HIV is notorious for its fast mutation and recombination. The identification of emerging genotypes, due to HIV mutation and recombination, presents a major challenge for the development of HIV vaccines and anti-HIV medicines [2]. In the last two decades, many different genotypes of HIV-1 have been reported, largely consisting of three major groups: M, O and N [3]. Further analyses have characterized the group M of HIV-1 into 9 subtypes (A–D, F–H, J, and K), and at least 16 circulating recombinant forms (CRFs) (`http://hiv-web.lanl.gov`). These subtypes represent different lineages of HIV-1, and have some geographical associations (Figure 1). Improved genotyping information will not only enhance the development of anti-HIV drugs and HIV vaccine, but also help us understand the epidemics of HIV infection. For example, previously placed subtypes E and I were later discovered to be recombinants. So it is clear that an efficient and effective genotyping system for HIV will be essential for HIV study.

Currently however, most genotyping methods are complicated and laborious. For a genotyping process that uses HIV-1 whole genomic sequences, there are two main challenges: (1) The mutation rates be-

*To whom correspondence should be addressed.

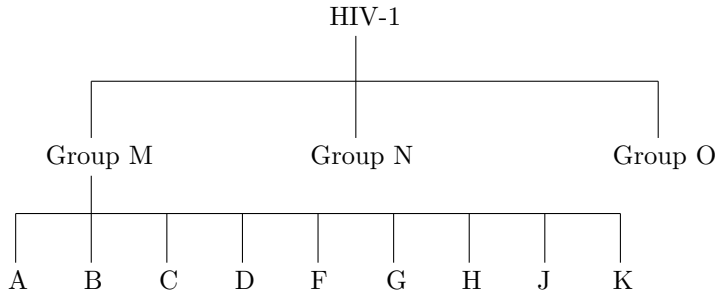


Fig. 1. This diagram illustrates the different levels of HIV-1 classification. HIV-1 is divided into three groups, and group M is divided into 9 subtypes.

tween HIV-1 genomes are not equal. For example, the genetic distance between genotypes is found to be greater in more polymorphic genes (e.g., the envelope gene) than in others. As a result, a phylogenetic relation based on a single gene may not be accurate with respect to the evolutionary patterns of HIV-1. Some recent methods propose the use of partial genome or whole genome sequences to conduct the phylogenetic analysis [4]. It is expected that recombination will complicate these genotype interpretation systems. (2) As the number of HIV-1 genomes increases, genotyping methods must be faster and more robust. Almost all of current available traditional genotyping methods are based on multiple sequence alignments [2, 5, 4]. For example, most of them rely on at least 300–500 character long alignment, which can then provide enough sequence information, even though the aligned length could vary for different regions of a genome. The importance of alignment length was emphasized in RevML [6], where it is reported that alignments with fewer than 400 characters generated trees with problems, such as mixing of sub-subtypes between A1 and A2, F1 and F2 and sometimes K, as well as mixing of B and D. However, we believe that the performance of multiple sequence alignments will decrease as the number of sequences increases. There are a number of genotyping systems especially designed for anti-retrovirus drug resistance studies [7, 2]. These genotyping methods either employ rule-based algorithms or are based on a genotype-phenotype database. However, these methods may generate some confusing instead of confirmatory information, because of the updating sequence information in the database [2].

In this paper we propose a novel method called the *whole genome composition distance* (WGCD), that avoids multiple sequence alignment to measure the evolutionary distance between two HIV-1 whole genomic sequences. Subsequently, a distance-based phylogenetic construction method Neighbor-Joining (NJ) [8] is adapted to build the phylogenetic clades. Our results show that the proposed approach can efficiently construct the phylogenetic clades for a set of 42 HIV-1 strains exactly the same as the one published by Los Alamos National Laboratory (LANL) with intensive computation and human curation [6].

The rest of the paper is organized as follows: In Section 2, we introduce in detail the whole genome composition distance computation using complete genomic sequences. Section 3 outlines the flow of operations in this novel HIV-1 genotyping system. An HIV-1 dataset that includes in total 42 whole genomic sequences is introduced in Section 4, as are the experimental results and our discussion. Experimental results and discussion on the individual CRF identification are also included in Section 4. We conclude the paper in Section 5.

2. WHOLE GENOME COMPOSITION DISTANCE

There are several existing HIV-1 genotyping systems, including the Stanford HIV-Seq program (`http://hivdb.Stanford.edu`), the NCBI Genotyping Program (`http://www.ncbi.nih.gov/projects/genotyping`), the Los Alamos Recombinant Identification Program (`http://hivweb.lanl.gov/RIP/RIPsubmit.html`), the European-based Subtype Analyzer Program (`http://pgv19.virol.ucl.ac.uk/download/star-linux.tar`) [9], and a recently developed system (`http://www.bioafrica.net/subtypetool/html`) described in [4]. All these systems employ a computationally intensive phase

of multiple sequence alignments or alike to align the query sequences, with priority given to some fragments that are known to be more conserved than the others. Consequently, when limited to single genes in the HIV-1 strains, these systems all perform acceptably well. However, the performance is highly dependent on the accumulated knowledge on HIV-1 strains, such as the fact that HIV-1 *pol* gene is highly conservative so at most two gaps can be introduced into the alignment.

On the other hand, different levels of conservation within different HIV-1 genes pose additional difficulties in the phylogenetic analysis: The analyses using different genes might produce inconsistent and even erroneous results (the same situation happens in numerous HIV databases) [10], and the multiple sequence alignments are more challenging, and thus the MSA-based phylogenetic analysis using multiple genes is computationally impossible.

In this study, we explore the possibility of HIV-1 phylogenetic analysis using their complete genomic sequences, through avoiding the computationally intensive phase of multiple sequence alignments. We adapt some ideas for whole genome phylogeny construction from the literature [11, 12] and propose a novel composition distance to measure the evolutionary closeness between two HIV-1 whole genomic sequences. After the distance matrix on the set of HIV-1 whole genomic sequences is thus computed, the Neighbor-Joining method [8] is utilized to build the phylogenetic clades. We note that there is a rich literature on whole genome phylogenetic analysis, where many approaches have been proposed for estimating the pairwise distance between two whole genomes. To name a few, there are approaches based on string composition [11–14], approaches based on text compression [15–18], and approaches based on gene content [19–22].

We use the whole genomic RNA sequences of HIV-1 to validate our method. Given an RNA sequence R, in our case to represent an HIV-1 strain, the *single nucleotide composition* of R is a vector of 4 nucleotide frequencies in sequence R. Namely, for each type of nucleotide nu, the frequency of nu in R is the number of occurrences of nu in sequence R divided by the total number of nucleotides in R. Likewise, the *dinucleotide composition* of R is a vector of $4^2 = 16$ dinucleotide frequencies in sequence R. In general, for each length-k nucleotide segment (there are possibly 4^k of them), its frequency in R is calculated as the number of its occurrences in sequence R divided by the total number of overlapping length-k nucleotide segments in R. For simplicity, the length-k nucleotide segment composition of R is called the k-th *composition* of R and denoted as $C_k(R)$.

The k-th composition of R can be used as a signature for strain R. Given two strains R and S, one may define, for example, the Euclidean distance between $C_k(R)$ and $C_k(S)$, each is a 4^k-dimensional vector, to measure the evolutionary distance between R and S. There are several similar measures proposed in the literature for whole genome phylogenetic analysis, including the single amino acid composition, the dipeptide composition [11], an SVD-based measure using tripeptide (and tetrapeptide) composition [13, 23], the complete information set (CIS) [12], and the composition vector (CV) [14]. In fact, the CIS method defines an information discrepancy between $C_k(R)$ and $C_k(S)$, and uses its normalized version to measure the evolutionary distance between R and S [12]. Based on our previous research on whole genome phylogenetic analysis for Avian Influenza Viruses (AIV), we propose to use the Euclidean distance between $C_k(R)$ and $C_k(S)$ to measure the evolutionary distance. That is, assuming that $C_k(R) = (f_1, f_2, \ldots, f_{4^k})$ and $C_k(S) = (g_1, g_2, \ldots, g_{4^k})$, where f_i and g_i are the frequencies of a common length-k nucleotide segment in R and S, respectively, then the Euclidean distance $d_k(R, S)$ is

$$d_k(R,S) = \sqrt{\sum_{i=1}^{4^k} (f_i - g_i)^2}. \qquad (1)$$

Note that in general $C_k(R)$ and $C_{k+1}(R)$ both contain evolutionary information of R, but some information hidden in one composition isn't necessarily revealed by the other. Obviously, the single nucleotide composition is one of the most information-rich compositions. The dinucleotide composition reveals the single nucleotide composition and some amount of extra evolutionary information not included in the single nucleotide composition. Likewise, the $(k+1)$-th composition is expected to contain some additional evolutionary information not included in the k-th composition. Nonetheless, this additional information is expected to decrease with increasing k (see also the Experimental Results and Discussion). For these reasons, we propose to use

$(C_1(R), C_2(R), \ldots, C_k(R))$, for a sufficiently large k, to represent strain R. Subsequently, assuming strain S is represented as $(C_1(S), C_2(S), \ldots, C_k(S))$, then the Euclidean distance between R and S is $d(R, S)$, defined as

$$d(R, S) = \sqrt{\sum_{i=1}^{k} d_i^2(R, S)}, \quad (2)$$

where $d_i(R, S)$ is defined in Equation (1). Such a distance is called the *whole genome composition distance* (WGCD) between R and S.

3. WGCD-BASED HIV-1 GENOTYPING

The WGCD-Based HIV-1 Genotyping system can be used as a tool, in addition to existing systems, typically in the cases where the whole strains are available. In fact, the genotyping system described here has been developed for analyzing whole strains to avoid the computationally intensive phase of multiple sequence alignments. The method can be roughly partitioned into the following four steps:

In the first step, a fixed nucleotide segment length k is used, which in our case is set to 80, to calculate the composition $(C_1(S), C_2(S), \ldots, C_k(S))$ for each HIV-1 whole strain S. Note that it is impossible to count the frequency for every length-80 nucleotide segment, as there would be $4^{80} \approx 1.5 \times 10^{48}$ of them. Instead, the counting is done for only those length-80 nucleotide segments that actually occur in S. This counting is computed by several linear scans of the whole strain S, and the frequencies of the same length nucleotide segments are alphabetically ordered. In the second step, for every pair of strains R and S, their evolutionary distance measured by $d(S, T)$ is computed using Equation (2). Note that the computation is done with a linear scanning the composition vectors for R and S at the same time, where the frequency for a non-occurring segment is automatically treated as 0. The resultant pairwise evolutionary distance matrix $M = (d(S, T))$ is fed to the Neighbor-Joining method in the third step to build a phylogenetic tree on the strains. The fourth step is the standard bootstrapping in phylogeny construction methods [24], to randomly mutate 30% of the genetic sequences at each iteration and then to build a bootstrapping tree using exactly the same method as stated in steps one to three. In total, 200 such bootstrapping trees are built and their consensus is produced as the final phylogenetic clades. Note that bootstrapping is used to test whether the output phylogeny is stable (or confident) with respect to the input sequences, given that there might be sequencing errors and some strains are incomplete sequences.

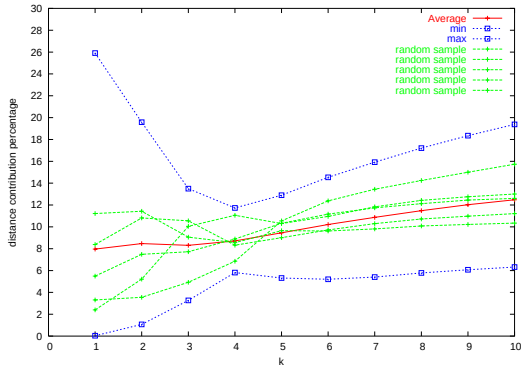
4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Datasets

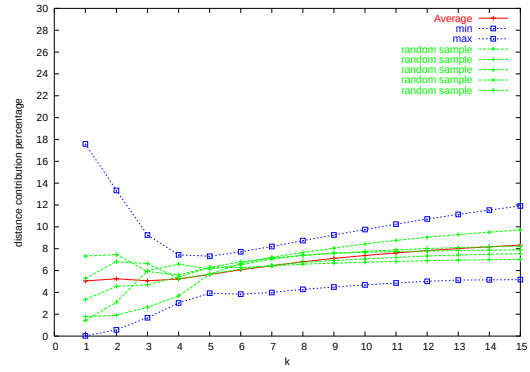
We downloaded a set of 42 HIV-1 strains based on a review paper [6], which were used as references sequences in genotyping using an enhanced version of fastDNAml maximum likelihood tree fitting (RevML) and site rate estimation codes (RevRates) [6]. The trees created in [6] require several executions of RevML and RevRates using different initial site mutation rates, followed by using modified global and local site mutation rates. Both trees on single genes and full-length *env*, *pol*, *gag* were constructed, and the single gene trees show problems when fewer than 400 character alignments occurred. In addition to these 42 HIV-1 strains, 2 CPZ strains (CPZ is a subtype of Simian immunodeficiency virus (SIV), which is believed to have a common ancestor to HIV-1 and HIV-2) were included for outgrouping purposes, as stated in the review paper [6]. The average length of these 44 whole genomic RNA sequences is $9,019$bp, with the maximum length of $9,829$bp and the minimum length of $8,349$bp. 15 recombinant strains were also downloaded for recombinant identification ability testing for the WGCD based HIV-1 genotyping system.

4.2. Results on Maximum Nucleotide Segment Length Determination

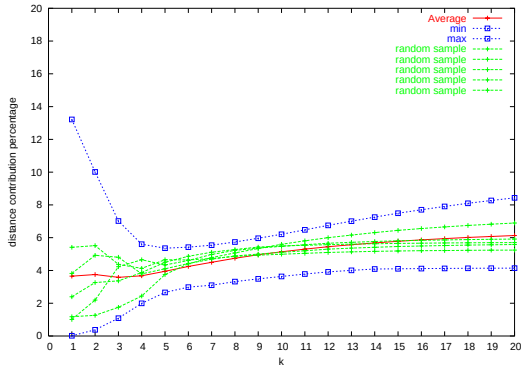
As discussed earlier, for any strain S, the $(k+1)$-th composition is expected to contain some additional evolutionary information not included in the k-th composition, yet this additional amount decreases with increasing k. We have set up experiments to validate this expectation, and subsequently to determine the maximum length k we should use in the WGCD-based genotyping. Given a pair of strains S and T, we calculate the composition vectors for them, respectively, i.e., $(C_1(S), C_2(S), \ldots, C_{80}(S))$



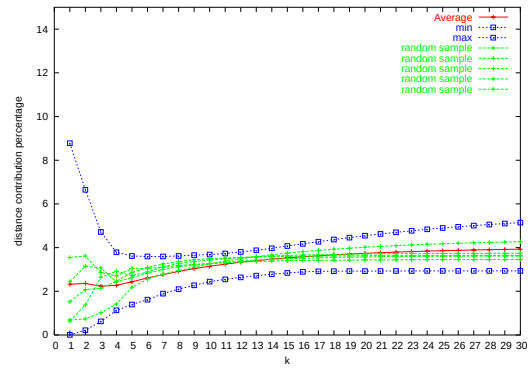
(a) $k = 10$.



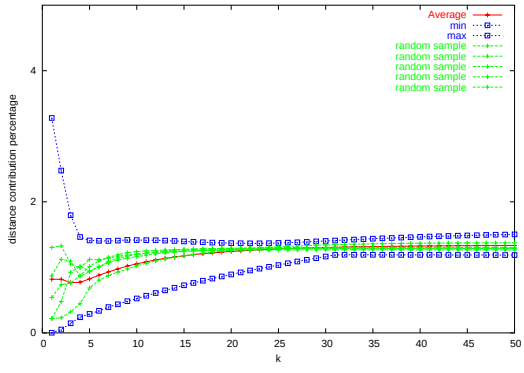
(b) $k = 15$.



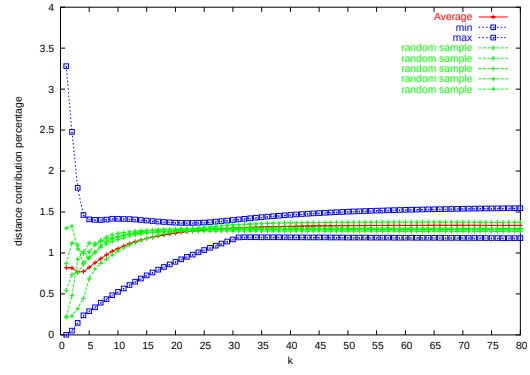
(c) $k = 20$.



(d) $k = 30$.

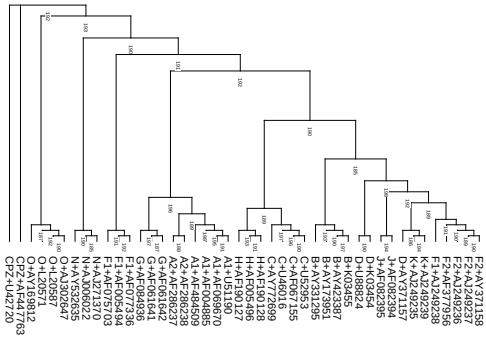


(e) $k = 50$.

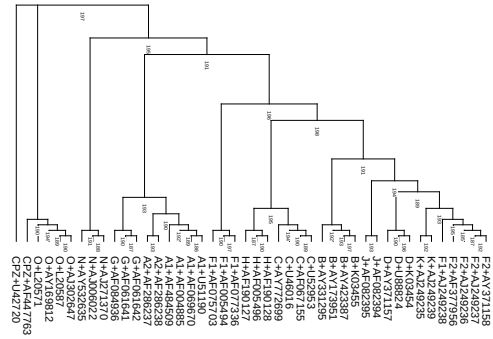


(f) $k = 80$.

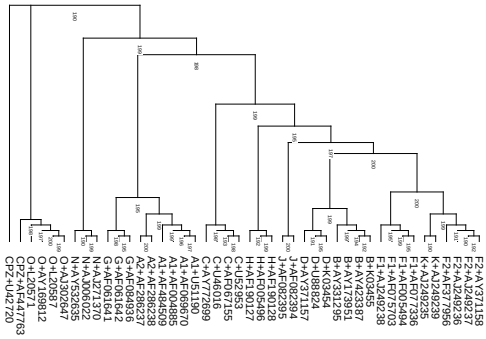
Fig. 2. The average distance contribution percentages of individual compositions to the WGCD with respect to different values of k, where green lines plot 5 randomly chosen distance contribution percentages and blue lines plot the minimum and maximum distance contribution percentages.



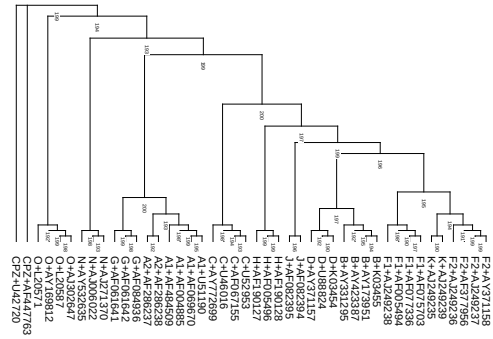
(a) $k = 10$.



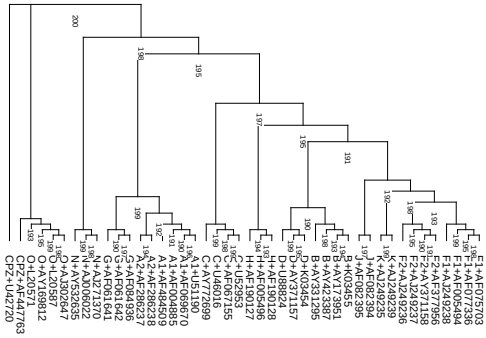
(b) $k = 15$.



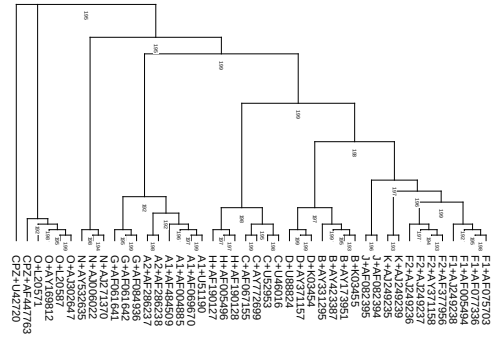
(c) $k = 20$.



(d) $k = 30$.



(e) $k = 50$.



(f) $k = 80$.

Fig. 3. The Neighbor-Joining phylogenies with 200 bootstrapping iterations on the WGCD using different values of k.

and $(C_1(T), C_2(T), \ldots, C_{80}(T))$. For each $i \leq 80$, we compute $d_i(S,T)$ according to Equation (1). We set $k = 10, 15, 20, 30, 50, 80$ to compute the WGCD between S and T, $d(S,T)$, using Equation (2). Subsequently, $c_i(S,T) = d_i^2(S,T)/d^2(S,T)$ is the *distance contribution percentage* of the i-th composition to the WGCD. We took the average of $c_i(S,T)$ over all pairs of strains and the average is denoted as $\overline{c_i}$, the average distance contribution percentage of the i-th composition to the WGCD. For $k = 10, 15, 20, 30, 50$ and 80, these $\overline{c_i}$'s are plotted as red lines in Figures 2(a)–2(f), where one can see that the tail portion of the line corresponding to $k = 80$ is approximately horizontal, but this is not the case for $k = 10, 15, 20, 30$, or 50. Note that we might think of volume $(\overline{c_{i+1}} - \overline{c_i})$ as the extra contribution of the $(i+1)$-th composition compared to the i-th composition. Therefore, the approximately horizontal tail portion indicates that we might neglect the evolutionary information carried by the nucleotide segments longer than 80.

4.3. Phylogenetic Analysis Results

We have constructed the phylogenies using $k = 10, 15, 20, 30, 50$ and 80. These trees are illustrated in Figures 3(a)–3(f), respectively. Comparing the contribution plots for individual compositions, the strains are classified better into the phylogenetic clades with increasing k. For example, in Figures 3(a) and 3(b), AJ249238 belongs to subtype F1, but is misclassified into F2; This problem is resolved in Figures 3(c) and 3(d), when k increases to 20 and 30, respectively. Nonetheless, subtype K is mis-inserted into subtype F in both of the phylogenies in Figures 3(c) and 3(d). When $k = 80$, a phylogeny which maps to all known evolutionary relationships is obtained (Figure 3(f)). For example, sub-subtypes A1 and A2 are adjacent to each other; sub-subtypes F1 and F2 are adjacent to each other; subtypes B and D are closer to each other than other subtypes; and groups M, N and O are well-separated.

For comparison, we have also conducted a control experiment to use the multiple sequence alignments by ClustalW (`http://www.ebi.ac.uk/clustalw/`). We uploaded all the 44 genomic sequences to the ClustalW webserver. The guide tree generated by ClustalW is shown in Figure 4. One can see that there is a problem with the outgrouping CPZ SIV strains, and subtype C is misplaced outside of group M.

We have also borrowed a software Biolayout [25, 26] (`http://www.biolayout.org/`) to display the phylogenetic clades. For this purpose, we removed the outgrouping CPZ SIV strains and sorted the pairwise distances between all 42 HIV-1 strains in increasing order. Scanning through the order, we selected the 84 minimum distances, which involve all 42 strains (84 is the minimum number such that this number of distances involve all 42 strains), and sent them to Biolayout. Figure 5 shows the graphical view of these distances, which clearly demonstrated 11 clades corresponding to 13 subtypes. Also can be seen are that subtypes A1 and A2, F1 and F2, B and D, D and G, and G and A2 are closer than the others.

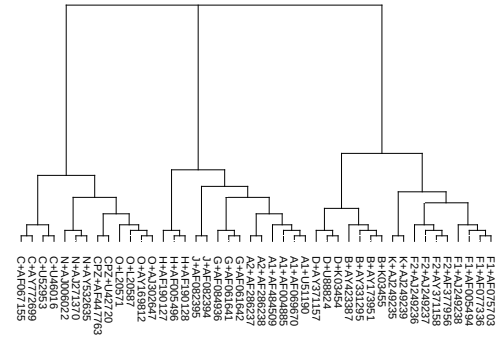


Fig. 4. The ClustalW guide tree on the 42 HIV-1 strains and two CPZ SIV strains.

4.4. Discussion

4.4.1. *Strains as Concatenations of Gene Sequences*

We have also used the concatenation of all 9 gene sequences to represent a strain, then used them to construct a phylogeny. For the set of the above 42 pure subtype strains and 2 CPZ strains, the resulting bootstrapping phylogenetic trees using $k = 80$ is shown in Figure 6. An interesting observation is that this phylogeny is different from the one constructed using complete genomic sequences — it contains several misplaced strains in subtypes A2 and C. In fact, this misplacement is consistent in all values of k we have tested, namely, $k = 10, 15, 20, 30, 50$ and

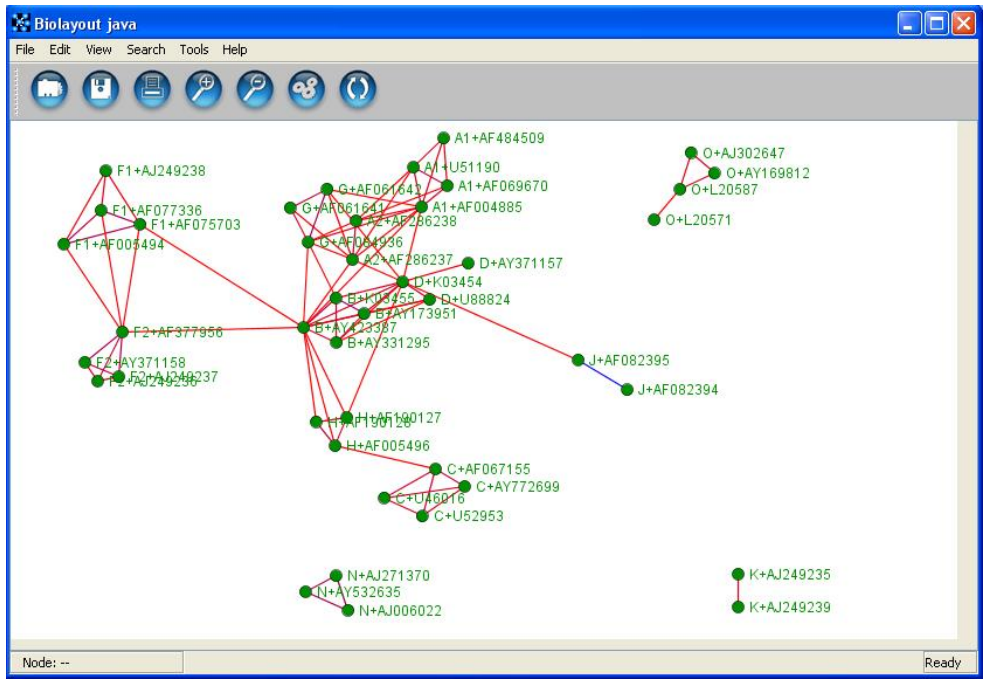


Fig. 5. The Biolayout graphical view of the phylogenetic clades of the 42 HIV-1 strains, using the 84 smallest distances computed by the WGCD method using nucleotide strings of length 1 to 80.

80. We have also obtained the distance contribution percentages of all the compositions, and the plots (as in Figure 2) are very "horizontal" when $k \geq 30$. This observation might indicate that the inter-gene regions in the complete genomic sequences for HIV-1 strains also contain evolutionary information that can be used for genotyping purpose.

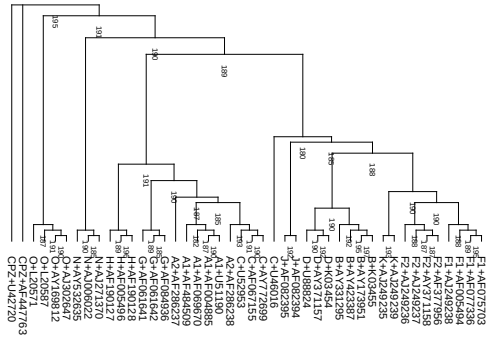


Fig. 6. The Neighbor-Joining phylogeny with 200 bootstrapping iterations on the WGCD using $k = 80$, where strains are represented as concatenations of 9 gene sequences each.

4.4.2. *Strains as Concatenations of Protein Sequences*

The gene products were also used for phylogenetic analysis. To this purpose, every strain is represented as a concatenation of 9 protein sequences. Since there are 20 types of amino acid residues, we set $k = 10, 20, 25$ and 40 in the WGCD distance computation. Once again, the distance contribution percentages were calculated and their plots (similarly as in Figure 2) are very "horizontal" when $k = 40$. The resultant bootstrapping Neighbor-Joining phylogeny for $k = 40$ is shown in Figure 7. In this phylogeny one can see that there are more misplaced phylogenetic clades than in the phylogeny using only gene sequences. For example, subtypes A1, C, F1, and G are split into 2 parts each; sub-subtypes F1 and F2, and sub-subtypes A1 and A2 become disconnected. Therefore, we might be able to conclude that using gene products for phylogenetic analysis on such fast evolving viruses is the least appropriate, and confirm that nucleotide sequences are superior. We note that this would be resulted from the silent mutations, which contribute to the genotyping diversity at nucleotide level but not protein level. In fact, the standard HIV-1 genotyping is done mostly

at the nucleotide level. Another possible reason that gene products are inferior may be that the evolution changes in the intergenic regions were not considered in the genotyping using protein coding regions.

4.4.3. *Identification of Circular Recombinant Forms*

Group M HIV-1 strains are currently classified into 9 subtypes and 16 circulating recombinant forms (CRFs). In our previous experiments, only non-CRF strains were included. Recently there are intensive activities in the area of discovering new HIV-1 CRFs, although many of the new CRFs have not yet been published [6] . In addition, evidence suggests many sub-subtypes and even a potential new subtype. The classification of new HIV-1 sequences follows the proposed HIV nomenclature guidelines [3, 27] to use the HIV-1 subtyping reference set. With respect to this set, each of the known CRFs was described by four representatives. It is recognized that, with the large increase of reported CRFs, the reference set would increase to an extent that it would cause problems in some analyses if four sequences were included for each CRF. Therefore, previous studies have limited the CRF section to one sequence per CRF, and the sequence selected for each subtype was intended to show how it is composed of the included subtypes.

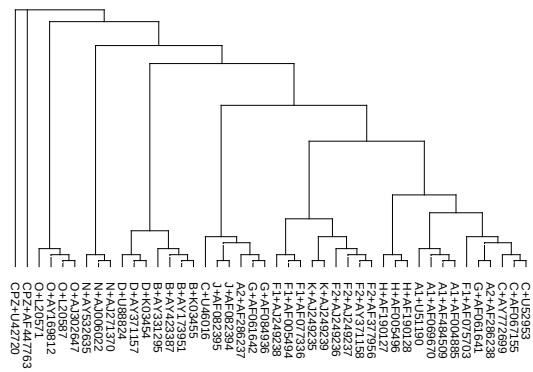
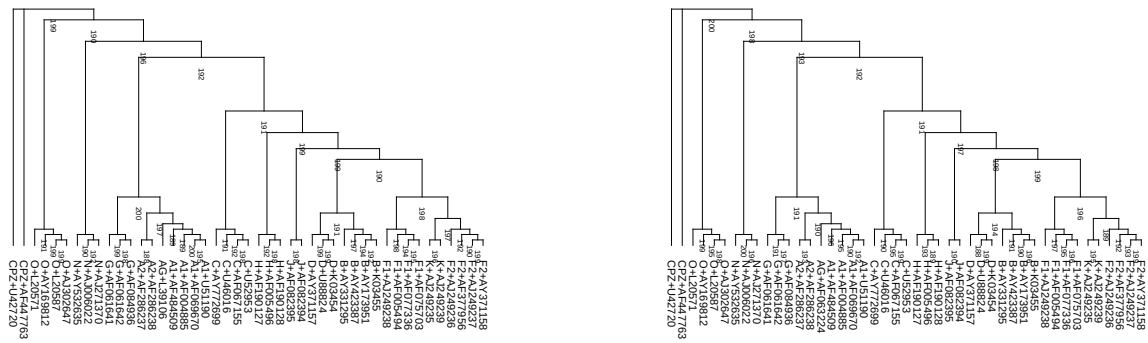


Fig. 7. The Neighbor-Joining phylogeny with 200 bootstrapping iterations on the WGCD using $k = 40$, where strains are represented as concatenations of 9 protein sequences each.

We have also performed experiments to test whether or not the WGCD-based genotyping system would be able to recognize CRFs. In the experiments, we found that when the number of CRFs is limited to one to two at a time (where CRFs are represented as complete genomic RNA sequences), then the recombinant subtype information can always be correctly identified. For example, Figures 8(a) and 8(b) show that both L39106 (NCBI accession number, strain from Nigeria) and AF063224 (NCBI accession number, strain from Djibouti) are recognized as A/G recombinant, and both of them are closer to subtype A than to subtype G, conforming with previous studies on these two recombinants. When we added both L39106 and AF063224 into the phylogenetic analysis, we found (Figure 9) they were grouped together and were recognized as A/G recombinants (a little closer to subtype A than to subtype G). Similarly, we have tested for 51 other CRFs individually and as a whole set, and in each case their recombinant form(s) was correctly identified (data not shown). Typically, we have also used Biolayout to graphically display the phylogenetic closeness relationships of L39106 and AF063224 to the pure subtype strains. Again for this purpose, the outgrouping CPZ strains were removed from the dataset, which consists of 42 HIV-1 pure subtype strains and 2 CRFs, and a total of 92 smallest distances were used. Figure 10 clearly shows the pure subtype clades and that both L39106 and AF063224 have edges connected to all A1, A2, and G strains. This fact indicates that neither of L39106 and AF063224 can be classified into A1, or A2, or G pure subtypes, yet they both have close relationships with A1, A2, and G subtypes, an obvious indication of recombinants. Based on these results, we are confident in claiming that, in terms of single CRF detections, our WGCD-based HIV-1 genotyping system performs well and that is superior to existing methods in the sense that it requires no CRF section pre-selection, nor that one CRF has to be described by four representatives.

Compared with other recombination detection systems based on multiple sequence alignments, the current version of our method has the limitation that it cannot detect the breaking points of the HIV-1 genotyping. In the future, we are proposing to check the motifs contributing the most to the recombination, and thus find the breaking points within the recombination. We will study how to detect the re-



(a) L39016.

(b) AF063224.

Fig. 8. The Neighbor-Joining phylogenies with 200 bootstrapping iterations on the WGCD using $k = 80$, where both L39106 and AF063224 are recognized as A/G recombinants, and both of them are closer to subtype A than to subtype G.

combination interactively, which will be able to find the recombinations from original subtypes and even between recombinant genotypes.

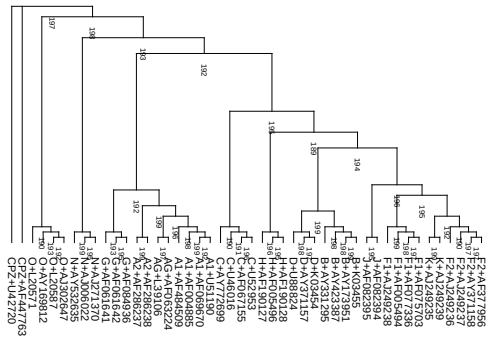


Fig. 9. The Neighbor-Joining phylogeny with 200 bootstrapping iterations on the WGCD using $k = 80$, where L39106 and AF063224 are grouped together and recognized as A/G recombinants (a little closer to subtype A than to subtype G).

4.4.4. *Subtype Signature Nucleotide Segment Identification*

In this current work, for a fixed k values, all nucleotide (or amino acid) segments of length up to k are used in the WGCD distance calculation. Looking at the resultant distance matrices, we have found that the pairwise distances are very close to each other, though they are sufficient in identifying the subtypes and CRFs. Therefore, one interesting research subject would be to identify those nucleotide segments that signify subtypes. With these signature segments, it could be easier for subtype identification for a newly sequenced strain, by simply looking at the occurrences of signature segments. Furthermore, occurrences of signature segments associated with multiple subtypes are indications of CRFs.

This is one of our on-going research focuses and we have obtained promising partial results. The full set of detailed experimental results will be reported in a succeeding paper. We are also constructing a web server based on the WGCD for HIV-1 genotyping, which will be released to public upon passing an extensive test.

5. CONCLUSIONS

We proposed a whole genome composition distance based HIV-1 genotyping system. Such a system avoids the computationally intensive phase of multiple sequence alignments, by using nucleotide segment frequencies to represent the genomes for phylogenetic analysis. Experiments show that every composition has its unique contributions to the evolutionary distance computation, and the contribution decreases with the nucleotide segment length increases. By

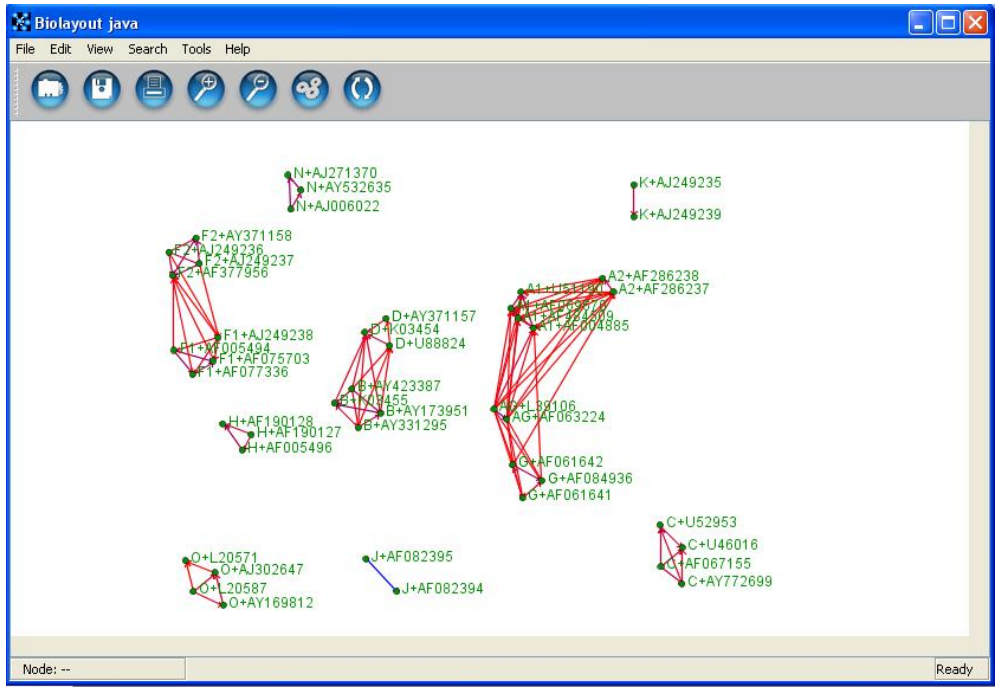


Fig. 10. The Biolayout graphical view of the phylogenetic clades of the 42 HIV-1 strains and 2 A/G CRFs, using the 92 smallest distances computed by the WGCD method using nucleotide strings of length 1 to 80. The figure shows that subtypes A (including A1 and A2) and G are fully connected by these two A/G CRFs.

setting the maximum segment length k to 80, the Neighbor-Joining method using the WGCD-based distances constructs phylogenetic clades that are identical to the golden standard ones determined by running several MSA-based genotyping systems together with manual parameter adjustments. Experiments on single CRF identification also confirm that such a new system is also capable of CRF discovery using only the whole strains, without any extra requirements on the CRFs. Lastly, our experiments also confirm that, for phylogenetic analysis on fast evolving viruses such as HIV, complete genomic sequences could be a better source than gene sequences and gene products.

ACKNOWLEDGMENTS

This research is supported in part by AICML, CFI and NSERC.

References

1. S. J. Popper, A. D. Sarr, K. U. Travers, A. Gueye-Ndiaye, S. Mboup, M. E. Essex, and P. J. Kanki. Lower human immunodeficiency virus (HIV) type 2 viral load reflects the difference in pathogenicity of HIV-1 and HIV-2. *Journal of Infectious Diseases*, 180:1116–1121, 1999.
2. M. Sturmer, H. W. Doerr, and W. Preiser. Variety of interpretation systems for human immunodeficiency virus type 1 genotyping: Confirmatory information or additional confusion? *Current Drug Targets Infectious Disorder*, 3:373–382, 2003.
3. D. L. Robertson, J. P. Anderson, J. A. Bradac, J. K. Carr, B. Foley, R. K. Funkhouser, F. Gao, B. H. Hahn, M. L. Kalish, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, P. M. Sharp, S. Wolinsky, and B. Korber. HIV-1 nomenclature proposal. *Science*, 288:55–56, 2000.
4. T. de Oliveira, K. Deforche, S. Cassol, M. Salminen, D. Paraskevis, C. Seebregts, J. Snoeck, E. J. van Rensburg, A. M. J. Wensing, D. A. van de Vijver, C. A. Boucher, R. Camacho, and A.-M. Vandamme. An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, 21:3797–3800, 2005.
5. M. Rozanov, U. Plikat, C. Chappey, A. Kochergin, and T. Tatusova. A web-based genotyping resource for viral sequences. *Nucleic Acids Research*, 32:W654–W659, 2004.
6. T. Leitner, B. Korber, M. Daniels, C. Calef, and B. Foley. *HIV-1 Subtype and Circulating Recom-*

binant Form (CRF) Reference Sequences. Accessible through `http://www.hiv.lanl.gov/content/hiv-db/REVIEWS/RefSeqs2005/RefSeqs05.ht%ml`, 2005.

7. R. W. Shafer, P. Hsu, A. K. Patick, C. Craig, and V. Brendel. Identification of biased amino acid substitution patterns in human immunodeficiency virus type 1 isolates from patients treated with protease inhibitors. *Journal of Virology*, 73:6197–6202, 1999.
8. N. Saitou and M. Nei. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4:406–425, 1987.
9. R. E. Myers, C. V. Gale, A. Harrison, Y. Takeuchi, and P. Kellam. A statistical model for HIV-1 sequence classification using the subtype analyser (STAR). *Bioinformatics*, 21:3535–3540, 2005.
10. The HIV Sequence Database. Accessible through `http://www.hiv.lanl.gov/content/hiv-db/mainpage.html`.
11. S. Karlin and C. Burge. Dinucleotide relative abundance extremes: a genomic signature. *Trends in Genetics*, 11:283–290, 1995.
12. W. Li, W. Fang, L. Ling, J. Wang, Z. Xuan, and R. Chen. Phylogeny based on whole genome as inferred from complete information set analysis. *Journal of Biological Physics*, 28:439–447, 2002.
13. G. Stuart, K. Moffet, and S. Baker. Integrated gene and species phylogenies from unaligned whole genome sequence. *Bioinformatics*, 18:100–108, 2002.
14. B. Hao and J. Qi. Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. In *Proceedings of the 2003 IEEE Bioinformatics Conference (CSB 2003)*, pages 375–385, 2003.
15. S. Grumbach and F. Tahi. Compression of DNA sequences. *Data Compression Conference*, 1993.
16. E. Rivals, M. Dauchet, J. Delahaye, and O. Delgrange. Compression and genetic sequences analysis. *Biochimie*, 78:315–322, 1996.
17. X. Chen, S. Kwong, and M. Li. A compression algorithm for DNA sequences and its applications in genome comparison. In *Proceedings of the Sixth Annual International Computing and Combinatorics Conference (RECOMB)*, pages 107–117. ACM Press, 2000.
18. D. Benedetto, E. Caglioti, and V. Loreto. Language trees and zipping. *Physical Review Letters*, 88:048702, 2002.
19. B. Snel, P. Bork, and M. A. Huynen. Genome phylogeny based on gene content. *National Genetics*, 21:108–110, 1999.
20. E. Herniou, T. Luque, X. Chen, J. Vlak, D. Winstanley, J. Cory, and D. O'Reilly. Use of whole genome sequence data to infer baculovirus phylogeny. *Journal of Virology*, 75:8117–8126, 2001.
21. C. House and S. Fitz-Gibbon. Using homolog groups to create a whole-genomic tree of free-living organisms: An update. *Molecular Evolution*, 54:539–547, 2002.
22. B. Snel, P. Bork, and M. A. Huynen. Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research*, 12:17–25, 2002.
23. G. Stuart, K. Moffet, and J. Leader. A comprehensive vertebrate phylogeny using vector representation of protein sequences from whole genomes. *Molecular Biology and Evolution*, 19:554–562, 2002.
24. J. Felsenstein. *PHYLIP*. Accessible through `http://evolution.genetics.washington.edu/phylip.html`.
25. An automatic graph layout algorithm for similarity and network visualization. *Bioinformatics*, 17:853–854, 2001.
26. L. Goldovsky, I. Cases, A. J. Enright, and C. A. Ouzounis. An automatic graph layout algorithm for similarity and network visualization. *Applied Bioinformatics*, 4:71–74, 2005.
27. D. L. Robertson, J. P. Anderson, J. A. Bradac, J. K. Carr, R. K. Funkhouser, F. Gao, B. H. Hahn, C. Kuiken, G. H. Learn, T. Leitner, F. McCutchan, S. Osmanov, M. Peeters, D. Pieniazek, M. Salminen, S. Wolinsky, and B. Korber. HIV-1 nomenclature proposal: a reference guide to HIV-1 classification. In *Human Retroviruses and AIDS 1999: a compilation and analysis of nucleic acid and amino acid sequences*. Los Alamos National Laboratory, Los Alamos, NM, 2000.