BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 14 2007, pages 1744–1752
doi:10.1093/bioinformatics/btm248

Phylogenetics

Nucleotide composition string selection in HIV-1 subtyping using whole genomes

Xiaomeng Wu[1], Zhipeng Cai[1], Xiu-Feng Wan[2], Tin Hoang[1], Randy Goebel[1] and Guohui Lin[1,*]

[1]Department of Computing Science, University of Alberta, Edmonton, Alberta T6G 2E8, Canada and [2]Department of Microbiology, Miami University, Oxford, OH 45056, USA

Received on November 11, 2006; revised on May 1, 2007; accepted on May 2, 2007

Advance Access publication May 11, 2007

Associate Editor: Joaquin Dopazo

ABSTRACT

Motivation: The availability of the whole genomic sequences of HIV-1 viruses provides an excellent resource for studying the HIV-1 phylogenies using all the genetic materials. However, such huge volumes of data create computational challenges in both memory consumption and CPU usage.

Results: We propose the complete composition vector representation for an HIV-1 strain, and a string scoring method to extract the nucleotide composition strings that contain the richest evolutionary information for phylogenetic analysis. In this way, a large-scale whole genome phylogenetic analysis for thousands of strains can be done both efficiently and effectively. By using 42 carefully curated strains as references, we apply our method to subtype 1156 HIV-1 strains (10.5 million nucleotides in total), which include 825 pure subtype strains and 331 recombinants. Our results show that our nucleotide composition string selection scheme is computationally efficient, and is able to define both pure subtypes and recombinant forms for HIV-1 strains using the 5000 top ranked nucleotide strings.

Availability: The Java executable and the HIV-1 datasets are accessible through 'http://www.cs.ualberta.ca/~ghlin/src/WebTools/hiv.php

Contact: ghlin@cs.ualberta.ca

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

The increased volume of available genomic data has made possible phylogenetic analysis for large sets of organisms at the whole genome scale. However, given that most genomes contain millions to billions of nucleotides, traditional molecular phylogenetic analysis approaches based on multiple sequence alignments, such as maximum parsimony and maximum likelihood, become impractical due to their high computational complexity. Moreover, different genes have different evolutionary rates; it has been shown that phylogenetic analyses using different (sets of) genes may give inconsistent results. For instance, for the *human immunodeficiency virus* (HIV-1), the envelope gene is known to evolve much faster than other genes (Leitner *et al.*, 2005); and for the *Ecdysozoa* clade of animals, the accepted reliability of 18S rRNA as a phylogenetic marker has been questioned (Dopazo *et al.*, 2004). Consequently, it is believed that sophisticated analyses on the whole genome sequences are required to provide a detailed and accurate picture of evolutionary relationships. However, the huge size of these whole genome sequences generally creates computational challenges including memory consumption and CPU usage.

There exist several attempts to address phylogenetic questions from a whole genome perspective, based on efficient information representation of the whole genomes while bypassing the high computational complexity stage of multiple sequence alignments (Chen *et al.*, 2000; Grumbach and Tahi, 1994; Hao and Qi, 2003; Herniou *et al.*, 2001; Karlin and Burge 1995; Milosavljevic, 1993; Rivals *et al.*, 1996; Snel *et al.*, 1999; Stauart and Berry, 2003; Stuart *et al.*, 2002a, b, 2004). All of these approaches are intended to extract the hidden evolutionary information from the whole genomes, but from different angles. For example, gene content based methods (Herniou *et al.*, 2001; House and Fitz-Gibbon, 2002; Snel *et al.*, 1999, 2000) mainly concentrate on a portion of homologous genes shared by multiple genomes, and then define an evolutionary distance between two genomes based on their gene sharing percentage. Alternatively, the compression based methods (Chen *et al.*, 2000; Grumbach and Tahi,1994; Milosavljevie,1993) generally regard the whole genomes as plain text, and define the similarity between two genomes as the relative compression ratio. The disadvantages of the above two approaches are first, that the former requires prior knowledge on homologous genes and second, the latter suffers from aggregate errors arising from compression.

The third class of methods in the whole genome phylogenetic analysis attempt to extend single nucleotide or single amino acid composition to study string composition for whole genomes where a string is a consecutive segment of nucleotides or amino acids (Hao and Qi, 2003; Karlin and Burge, 1995; Li *et al.*, 2002; Qi *et al.*, 2004; Stuart and Berry, 2003; Stuart *et al.*, 2002a, b, 2004). Recent proposals in this category include Karlin and Burge (1995), who analyzed the systematic

*To whom correspondence should be addressed.

differences in dinucleotide frequencies within and between species, and obtained a biologically plausible phylogenetic tree for mitochondrial genomes, Hao and Qi, (2003); Li *et al*., (2002); Qi *et al*., (2004), who analyzed asymmetries in length-k word distribution, then extracted phylogenetic properties from genome-wide statistical observables for prokaryotes, and Stuart and Berry , (2003); Stuart *et al*., (2002a, b, 2004), who used singular value decomposition (SVD) to analyze short peptide frequencies (of length 3–5), then built species phylogenies.

In the reported experimental results, all of the above mentioned methods in the third category managed only strings of length 7 or less for amino acid sequences and of length 12 or less for nucleotide sequences into computation, because of memory demands. Theoretically, one may increase the maximum string length to have finer composition for the whole genomes in order to obtain more accurate pair-wise evolutionary distances. However, increasing string length requires too much memory to be practical as well as increased CPU usage. For example, computing length-7 peptide composition for a whole genome (which is regarded as the union of its encoding proteins) already requires gigabytes of storage, regardless of the size of the genome. Consequently, in practice, the maximum string lengths have been set to relatively small values such as 5 and 6.

Nevertheless, it has also been observed that not every composition string contributes equally to the evolutionary distance calculation. In fact, some strings appear to have more discriminatory power than others. Stuart and Berry, (2003); Stuart *et al*., (2002a, b, 2004) proposed to employ SVD on the peptide-to-genome frequency matrix, to extract a reduced number of string linear combinations, and then use their pseudo frequencies to represent the genomes. However, such a decomposition does not address the memory issue, i.e. the peptide-to-genome frequency matrix must still be computed, and that can only be done when the maximum string length is a small value. In addition, the string linear combinations created by SVD are difficult to explain biologically.

Based on the above two major observations, we propose a string scoring method to extract explicit composition strings that heuristically identify the richest evolutionary information, and then to use only these selected strings in the evolutionary distance calculations. These selected composition strings can be regarded as the most important features with respect to the whole genomic sequences. In our method, the number of selected strings is a parameter that can be tuned depending on the available computing resources. In particular, the memory requirement in the selection process is proportional to the number of selected strings, and selecting thousands of strings can be processed on a normal desktop, while examining candidate strings of an arbitrarily large length.

We applied our method on a dataset of 867 pure subtype HIV-1 strains and 331 various recombinants, to predict their subtypes or recombinant forms. Among the 867 pure subtype strains, 42 were used as references. By setting the number of selected strings at 500 and the maximum string length at 21, we achieved 100% leave-one-out subtyping accuracy on the reference dataset of 42 pure subtype strains. Using these 500 strings, we also achieved 100% subtyping accuracy on the independent testing dataset of the other 825 pure subtype strains. These 867 pure subtype strains were also used in blind comparison to three most recently proposed HIV-1 subtyping programs (Myers *et al*., 2005; Oliveira *et al*., 2005; Rozanov *et al*., 2004) which achieved 96.4, 99.2, and 99.5% accuracy, respectively. More detailed analysis revealed these 500 top scoring strings to be signature strings associated with certain subtypes. Subsequently, we present a method to remove 2–50% consecutive nucleotides from each of the 331 recombinant strains and then to predict the subtype information for the remaining sequence. The non-trivial percentage (e.g. 3%) of predicted subtypes match well with the known recombinant forms, with some exceptions strongly suggesting the need for further human re-curation. All these results demonstrate that our proposal is promising in terms of both the biological significance of the selected nucleotide composition strings and the quality of the recovered phylogenetic relationships.

The rest of the article is organized as follows: in the next section, we briefly introduce the *Complete Composition Vector* (CCV) representation for a whole genome. We will then present a selection scheme to extract the most informative nucleotide composition strings. Using these selected strings, we can obtain a much lower dimensional composition vector for each genome. We then define the evolutionary distance between two genomes based on their composition vectors. In Sections 3 and 4, we report and discuss the computational results on the HIV-1 subtyping, respectively. Section 5 presents the recombinant form prediction and the preliminary experimental results. We conclude the article in Section 6.

2 METHODS

2.1 Complete composition vector

We use whole genomic sequences to introduce the concept of CCV. For a genome represented as the union of its encoding protein sequences, its CCV can be analogously defined. First, a length-k string is a sequence of k consecutive nucleotides. Given a whole genomic sequence G of length L, the number of appearances of a length-k string $\alpha = a_1 a_2 \dots a_k$ in G is $f(\alpha)$, where every a_i is a nucleotide. Since there are $L-k+1$ (overlapping) length-k strings in G in total, the *probability* of appearance of string α in sequence G is $p(\alpha) = f(\alpha)/(L-k+1)$. Similarly, we can define the probability of appearance $p(\alpha)$ for string α in a whole genome containing multiple chromosomes, where the dividend becomes the number of appearances across all the chromosomes and the divisor becomes the total number of (overlapping) length-k strings in all the chromosomes.

Based on all the string appearance probabilities, we can define the composition value $\pi(\alpha)$ for string α. In this article, we adopt a second order Markov model similar to Hao and Qi (2003). In such a model, we first calculate the expected appearance probability of string $\alpha = a_1 a_2 \dots a_k$ as $q(a_1 a_2 \dots a_k) = (p(a_1 a_2 \dots a_{k-1}) \times p(a_2 a_3 \dots a_k))/p(a_2 a_3 \dots a_{k-1})$, and then define the composition value $\pi(\alpha) = (p(\alpha) - q(\alpha))/q(\alpha)$. All the composition values are stored in a sequential order to form a vector $V_k(G) = \langle \pi_1, \pi_2, \dots, \pi_m \rangle$ that represents the whole genome G, where k is the string length and m denotes the total number of strings under consideration. In Hao and Qi (2003) and Qi *et al*. (2004), the (amino acid) string length k was fixed at a single very small value (≤ 6). In one of our previous research (Wu *et al*., 2006), we conducted a systematic

study and concluded that using strings with multiple lengths, in a range $[1, K]$ for some K, is more effective. Particularly, the phylogenetic analysis and the resultant phylogeny in Wu *et al.* (2006) showed improvements over using only one fixed length. In this article, we continue to use strings of length range $[1, K]$, and the vector definition by all these composition values of strings, i.e. the concatenation of $V_1, V_2, \ldots, V_K$, is referred to as the CCV of the whole genome. Certainly, a larger value of K gives a vector containing finer evolutionary information.

A CCV is thus an m-dimensional vector (for instance, m could be as large as $4 + 4^2 + 4^3 + \ldots + 4^{15} = 1431,721,300$, when $K = 15$). Note that m could be a very large number, and it implies one major disadvantage of CCV for acquiring too much memory to be computationally efficient. In the preliminary experiments, we set target to examine strings of length up to 100, and therefore the memory issue needed to be addressed. First, observe that there are strings, especially when they are long, which do not occur at all in any whole genome in the dataset. We thus do not compute their composition values. Subsequently, the CCV for a whole genome has a much lower dimension, in which every entry is associated with a string that occurs in at least one genome. Second, notice that not all strings contribute to the phylogenetic analysis equally. Therefore, we propose to extract only a small number of strings, which contain the richest evolutionary information, and use only them in the phylogenetic analysis. The proposed string extraction scheme is based on the measurement of *relative entropy*, which has been constantly employed in the general feature selection in the statistical learning literature. The most important parameter in this framework is the number of extracted strings, which is likely dataset dependent and, on the HIV-1 subtyping, is set at 500 through extensive preliminary/training experiments. Such a setting is to maintain the overall quality of the recovered HIV-1 phylogenetic relationships. Under this string extraction scheme, we were able to examine much longer strings (in our experiments, up to 100) without causing any memory problems, and as a result we discovered that those composition strings with the richest evolutionary information in HIV-1 subtyping have length mostly in the range $[5, 9]$. Such a discovery confirms partially the idea that we can skip long strings in the whole genome phylogenetic analysis. The reported HIV-1 subtyping results are on strings of length 1–21. It also confirms that using a single length is not sufficient (Wu *et al.*, 2006), and thus the CCV representation is in general more effective than the representation proposed in Hao and Qi (2003) and Qi *et al.* (2004).

2.2 String selection and phylogenetic relationships

In this section, we first introduce a scoring scheme to estimate how important a nucleotide composition string is, and then, by selecting top ranked 500 strings, we obtain a 500-dimensional composition vector for each whole genome. Two other scoring schemes that have also been tested and the empirical determination of the string number 500 are included in Discussion section. On the HIV-1 dataset of 42 reference strains, we note that 500 is much smaller than the total number of examined strings (of length 1–21), which is 2 260 957, and these 500-dimensional vectors can be computed without causing any memory problem. Note also that, disregarding the memory issue, the string selection is done in almost the same amount of time for computing the CCV representation, except a negligible amount of time for computing relative entropies. These vectors are then used to define a pair-wise evolutionary distance between every pair of whole genomes, and then the achieved distance matrix is used to construct a Neighbor-Joining (Saitou and Nei, 1987) phylogeny, and in subtyping. The quality of the recovered phylogenetic relationships, represented in subtyping accuracy, demonstrates the success of our method and the quality of the selected composition strings.

2.3 String scoring scheme: relative entropy

Basically, the scoring scheme is set up to evaluate the information content associated with the composition strings, and to assign a higher score to a string if its information content is richer. Note that each string is evaluated independently. To begin, we concatenate all the given whole genomes in the dataset and regard the result as a *super-genome*. We then compute the composition value $\pi(\alpha, i)$ for string α in genome i, for each $i = 1, 2, \ldots, n$ (here n is the number of whole genomes in the dataset), and the composition value $\Pi(\alpha)$ for string α in the super-genome.

The absolute composition values $|\pi(\alpha, i)|$ for string α in all the given genomes may be regarded as an *unnormalized* probability distribution of string α, where the index i is regarded as a variable. We use *relative entropy* (or Kullback–Leibler distance) to assign a score to string α to measure the distance between this distribution and the *unnormalized* background probability represented as $\Pi(\alpha)$. Namely,

$$s(\alpha) = \sum_{i=1}^{n} |\pi(\alpha, i)| \ln \left| \frac{\pi(\alpha, i)}{\Pi(\alpha)} \right|,$$

where ln( ) is the natural logarithm. Note that relative entropy is used to estimate the distance between two probability distributions. Therefore, $s(\ )$ defined in the above is close to 0 if the actual distribution is close to the background one. In other words, the larger the absolute relative entropy, the more informative string α is.

2.4 Selected string composition vector

We maintain a buffer of size 500 to store the nucleotide composition strings that have been examined, and have the highest scores using the above relative entropy-based scoring scheme. We examine the strings in increasing length and, for each length, in lexicographical order. For each string under consideration, if there is a room in the buffer (i.e. among the first 500 strings), it is appended; otherwise, its score is compared with the minimum score of the strings stored in the buffer, and if larger, it replaces the string with the minimum score. By only saving these 500 highest scored strings, the potential memory issue is resolved and the maximum string length to be examined can be set to an arbitrarily large value. For example, we have examined strings of length 100 in our preliminary experiments on a normal desktop of 1 GB memory. After all strings have been examined, the composition values of the 500 top scored strings stored in the buffer are used to assemble the 500-dimensional composition vectors to represent the whole genomes. Let $V(i) = \langle \pi_{i_1}, \pi_{i_2}, \ldots, \pi_{i_m} \rangle$ be the vector representing genome i, for $i = 1, 2, \ldots, n$, where $m = 500$ and π_{i_j} denotes the composition value of the j-th highest scored string in genome i, for $j = 1, 2, \ldots, m$.

2.5 Pair-wise evolutionary distance

For every pair of genomes i and j, represented as vectors $V(i) = \langle \pi_{i_1}, \pi_{i_2}, \ldots, \pi_{i_m} \rangle$ and $V(j) = \langle \pi_{j_1}, \pi_{j_2}, \ldots, \pi_{j_m} \rangle$ in the m-dimensional space, the Euclidean distance between them is

$$d(i, j) = \left(\sum_{\ell=1}^{m} (\pi_{i_\ell} - \pi_{j_\ell})^2 \right)^{\frac{1}{2}},$$

which is taken as the evolutionary distance between these two genomes. This gives a distance matrix $D_{n \times n} = (d(i, j))_{n \times n}$ for the n genomes in the input dataset.

2.6 Whole genome phylogenetic relationships

The distance matrix $D_{n\times n}$ is used as input to the Neighbor-Joining algorithm to display the phylogenetic relationships among the whole genomes. These distances are also used for HIV-1 subtype prediction for each testing strain. Essentially, the distances between the testing strain and the carefully chosen 42 HIV-1 reference strains are calculated using the above steps of operations, and based on them the subtype or the recombinant form of the testing strain is inferred.

3 COMPUTATIONAL RESULTS

3.1 Overview

To evaluate the effectiveness of our string selection method, we have tested it on a dataset of HIV-1 pure subtype and recombinant strains to predict their subtypes or recombinant forms.

HIV is among the most genetically variable organisms known. HIV-1 is classified into three major phylogenetic groups, M (major), N (new) and O (others). Group M, which is responsible for the HIV pandemic, is further divided into nine subtypes, some of which have been even further subdivided into sub-subtypes. Besides GenBank, there are several other viral databases holding HIV virus sequences, such as the one provided by Los Alamos National Laboratory (http://hiv-web.lanl.gov/content/index). In 2005, a set of 42 reference whole genomic sequences was published (Leitner *et al*., 2005), which included 35 sequences in group M, 3 in group N and 4 in group O. In addition, we have downloaded a total of 825 other pure subtype and 331 recombinant HIV-1 whole genomes for independent testing (several hundred other incomplete strains were excluded from our experiments).

Accurate determination of the genetic subtype for an HIV-1 strain is of crucial importance for epidemiological monitoring as well as for the design of molecular detection systems and potential vaccines (Rambaut *et al*., 2004). This work addresses mainly the pure subtype HIV-1 subtyping. A discussion on using this subtyping system to determine the recombinant form for a recombinant strain is included in Section 5. Current subtyping and recombinant form determination methods mostly rely on multiple sequence alignments (de Oliveira *et al*., 2005; Martin *et al*., 2005; Myers *et al*., 2005), except the one by Rozanov *et al*. (2004) based on BLAST search (Altschul *et al*., 1997). To the best of our knowledge, multiple sequence alignments have limited quality and are constrained by the size of the dataset (e.g. the EMBL-EBI ClustalW server at 'http://www.ebi.ac.uk/clustalw/' accepts datasets containing no more than 500 sequences). On the set of 42 HIV-1 reference strains, we selected the 500 top scored strings by relative entropy, and the leave-one-out subtyping accuracy using only the 500 selected strings was 100%. Using these 500 strings, independent subtyping of the 825 testing strains achieved also 100% accuracy. The combined dataset of 867 pure subtype strains were also sent to three other HIV-1 subtyping programs (de Oliveira *et al*., 2005; Myers *et al*., 2005; Rozanov *et al*., 2004), for comparison purpose. Overall, our method successfully avoids the computationally intensive alignment phase, and achieves high subtyping accuracy. Another advantage of our method is that it does not require any pre-knowledge about the genomic sequences (such as one portion of the genome is more important than the other portion during the subtyping), while those important regions will be automatically detected according to their coverage by the selected strings, which also allow biological explanation.

3.2 Results

We applied our string selection method on the set of 42 HIV-1 reference sequences, which can be viewed as the training stage, to select the 500 most informative strings for subtyping purpose. The discerning power of these strings is evaluated through the leave-one-out cross-validation on the 42 reference sequences and an independent testing on the dataset containing 825 pure subtype HIV-1 viral sequences. The 42 HIV-1 reference sequences consists of 6 subtype A (4 A1 and 2 A2), 4 subtype B, 4 subtype C, 3 subtype D, 8 subtype F (4 F1 and 4 F2), 3 subtype G, 3 subtype H, 2 subtype J, 2 subtype K, 3 type N and 4 type O. The average length of these strains is 9005 bp, with the maximum length 9829 bp and the minimum length 8349 bp. These HIV-1 reference sequences were carefully selected by considering several criteria (Leitner *et al*., 2005).

We set a maximum string length K in our method, and the method examined the strings in increasing length and, for each length, in lexicographical order. When $K = 21$, the 500 top ranked (out of 2 260 957) strings have their length distributed in between 5 and 10. The second row of Table 1 shows the percentages of different length strings among these 500 top ranked strings, where one can see that length-7 and length-8 strings are dominant (81.0%). We have also collected the relative entropies of the top 5000 strings (whose percentages in the third row of Table 1) and plotted them in Figure 1 in non-increasing order. The top 500 strings are colored blue (the other 4500 in red) in 1. Form the plot, it is clear that the strings that did not make into the list of 500 have relatively small relative entropies, and thus can largely be ignored. By representing each strain as a 500-dimensional vector, the subsequently computed evolutionary distance matrix for this set of 42 HIV-1 reference strains was used as input to the Neighbor-Joining method to generate a phylogenetic tree (Fig. 2), using one CIV strain `AF447763` as an outgroup. In this tree, all subtypes are clearly grouped together as distinct branches, and the closeness relationships among the subtypes are also well demonstrated, e.g. subtypes B and D are closer to each other than to the others and subtype F (A) indeed contains two distinguishable sub-subtypes F1 and F2 (A1 and A2, respectively).

On these 42 HIV-1 reference sequences, we adopted the leave-one-out cross-validation (LOOCV) scheme to predict the subtype information for each sequence whose subtype was

Table 1. Percentages of different length strings in the top ranked strings

Length	4	5	6	7	8	9	10
Top 500	–	2.2	10.0	58.2	22.8	6.6	0.2
Top 5000	0.138	1.308	7.108	22.85	37.122	24.006	7.462

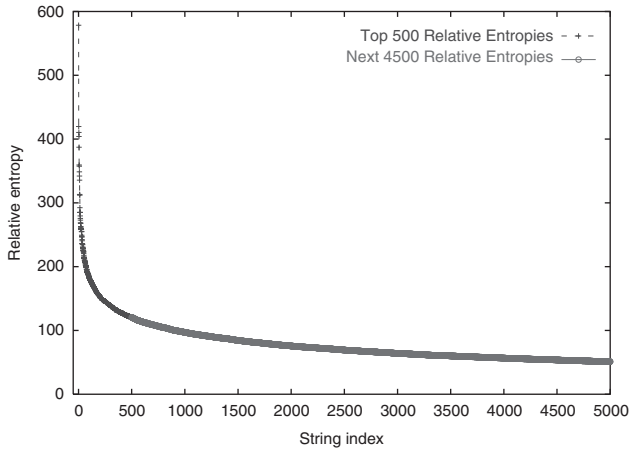

Fig. 1. The relative entropy scores of the 5000 top ranked strings, in decreasing order, in which the 500 top ones are colored blue. Colour version of this figure is available as Supplementary material online.

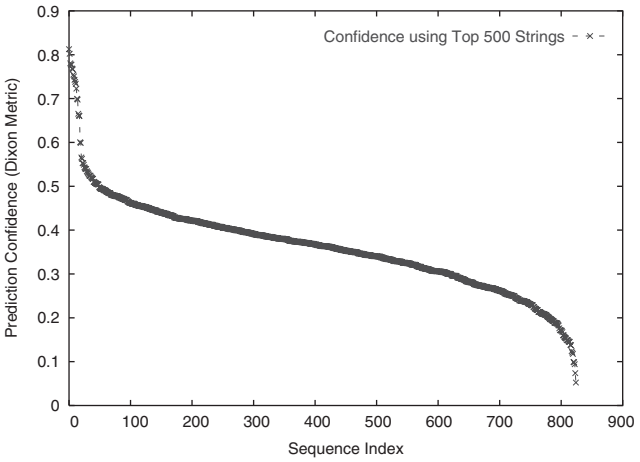

Fig. 3. Subtype prediction confidence values (Dixon metric) in non-increasing order, using the top 500 strings. Only 5 out of the 825 predictions are considered not-so-confident under Dixon metric. Colour version of this figure is available as Supplementary material online.

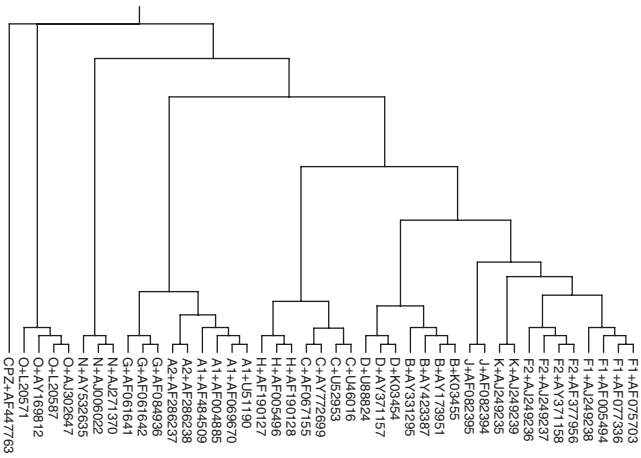

Fig. 2. The Neighbor-Joining phylogenetic tree on the 42 reference sequences using the 500 top ranked strings, one CIV strain AF447763 is used as an outgroup.

blinded. In more details, the testing sequence was removed from the dataset, and the above string selection procedure was applied to the rest of the 41 sequences to identify the 500 top ranked strings by their relative entropies. Note that these 500 strings could slightly differ from the 500 top ranked strings by using all 42 sequences. Next, using the selected 500 strings, the distances between the testing sequence and all the 41 reference sequences were calculated, and the subtype of the closest reference sequence was taken as the predicted subtype of the testing sequence. We repeated this training testing on each of the 42 sequences and obtained 100% subtyping accuracy.

Using these 42 reference sequences as a training dataset to select 500 strings, we independently predicted the subtype for each of the 825 pure subtype HIV-1 sequences. Among the 825 sequences, there are 55 A1, 9 A (not known to be A1 or A2), 264 B, 415 C, 51 D, 2 F1, 10 G, 2 N and 17 O. For each of the testing sequences, whose subtype was blinded, the distances between it and all the 42 reference sequences were calculated using (*only*) the 500 selected strings. The subtype of the closest reference sequence was then taken as the predicted subtype of the testing sequence. We also achieved 100% subtyping accuracy (for each of the 9 A sequences, both A1 and A2 were counted as correct prediction) on this independent testing dataset. Moreover, for each testing sequence, we have observed that the second closest reference sequence from the 42 reference sequences has the same subtype as the closest one. This certainly indicates high prediction confidence. For each testing sequence, we have also calculated its average distances to all the 13 subtypes. The closest subtype by average distance is exactly the same as the subtype of the closest sequence. Let d_1 and d_2 denote the shortest and the second shortest average distances, respectively, and d_{13} denote the longest average distance. Numerically, we assigned $(d_2 - d_1)/(d_{13} - d_1)$ (*Dixon metric*) as the quantified confidence associated with the subtype prediction. For all the 825 testing sequences, their subtype prediction confidences are plotted in Figure 3, in non-increasing order, where only 5 of them are less than 0.1 (0.099233, 0.098523, 0.094155, 0.073713 and 0.052269), which is the normal lower bound for high confidence (Su *et al*., 2001). A closer look at these five sequences tells that (1) four of them are of subtype D, and their average distances to subtypes D and B are very close to each other; (2) the other one (AY173955) is of subtype B, whose average distance to subtype D is very close to its average distance to subtype B.

To compare with existing HIV-1 subtyping programs (de Oliveira *et al*., 2005; Myers *et al*., 2005; Rozanov *et al*., 2004), we have uploaded all the 867 pure subtype sequences to them to predict their subtypes. The genotyping tool by Rozanov *et al*. (2004) (http://www.ncbi.nlm.nih.gov/projects/genotyping/) slides a window along the query sequence and BLASTs each window segment against reference sequences. Similarity scores to reference sequences are returned for each BLAST, and we applied the naive pure subtype assignment using the subtype of the reference sequence with the highest average similarity score. Its overall prediction accuracy was 99.5% (four were predicted incorrectly, precision 99.5%,

Supplementary Materials), but when we forced it to predict pure subtypes only, its accuracy reached 100%. The subtyping system, BioAfrica, by de Oliveira *et al.* (2005) (http://www.bioafrica.net/subtypetool/html/) consists of a multiple sequence alignment by ClustalW, maximum likelihood phylogenetic analysis by PAUP followed by bootscanning and subtype determination by Treepuzzle. Its prediction accuracy was 99.2% (seven were unassigned, no false positive, Supplementary Materials). The STAR subtyping system by Myers *et al.* (2005) (http://www.vgb.ucl.ac.uk/starn.shtml) evaluates the query sequence against subtype profiles and returns discrimination scores, which are then transformed into a Z-score distribution for determination of HIV-1 subtype. Its prediction accuracy was 96.4% (31 were unassigned, no false positive, Supplementary Materials). There is a recent independent assessment (Gifford *et al.*, 2006) of three automated genotyping tools including the above BioAfrica and STAR, which (was brought to our attention during the revision) shows many inconsistent genotyping results and suggests that those unassigned strains/sequences show some evidence of recombination. The 867 strains have been annotated as pure subtypes, and our method did perform well, though it might seem aggressive on the not-so-confident prediction.

4 DISCUSSION

We mentioned that the maximum string length K to be examined did not affect the subtyping performance in our experiments, as long as it is larger than a certain value C. On the set of 42 HIV-1 pure subtype strains, C is 10 when we set the number of strings to be selected to 5000 or less. Nevertheless, C is clearly associated with the number of selected strings, and a larger number of selected strings would imply a larger value of C. We also believe that C is dataset dependent. That is, for other whole genomic sequences, C could have different values, even when the number of selected strings is the same. We will be investigating this idea on avian influenza virus (AIV) and food and mouth disease virus (FMDV).

To address whether 500 top ranked strings by relative entropy are appropriate for HIV-1 subtyping, we examined a range of selected strings from 50 to 5000, at the increment of 50, and checked the corresponding subtyping accuracy. We have observed the first 100% subtyping accuracy at 450, and for every other number tested afterwards, the subtyping accuracy remained at 100% (i.e. never decreased). We therefore decided to set 500 as the default. Nevertheless, we have found that, by using 5000 strings, most of the subtyping confidences increased. In particular, four not-so-confident predictions using only 500 strings became confident when using 5000 strings, and only one (`DQ054367`) remained not-so-confident (the Dixon metric decreased, strangely, from 0.098523 to 0.080143). On top of the non-increasing order of prediction confidences using 500 strings (blue), the prediction confidences using 5000 strings (red) are plotted and shown in Figure 4, where one can see that the relative confidences remain largely unchanged and, for most of the testing sequences, the associated prediction confidences using 5000 strings increase.

Next, we examined how well the selected 500 strings cover the positions in the HIV-1 whole genomes. For each of these

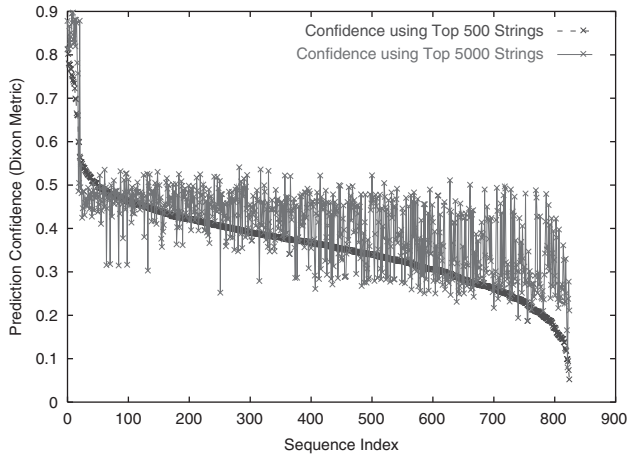

Fig. 4. The prediction confidence values using the top 5000 strings plotted on top of the order by using the top 500 strings. Only one prediction using the top 5000 strings remains not-so-confident. Colour version of this figure is available as Supplementary material online.

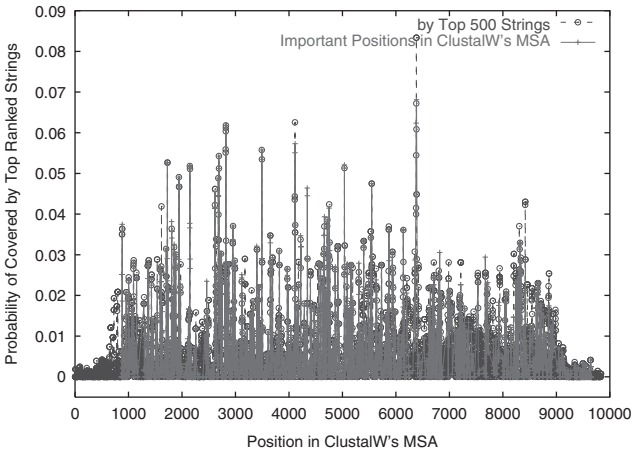

Fig. 5. ClustalW's MSA position coverage by the 500 top ranked strings, for the 42 reference sequences. Colour version of this figure is available as Supplementary material online.

500 selected strings, if it occurs in one of the 867 pure subtype sequences, then the positions where the string occurs receive a probability of $k/(L-k+1)$ each, where k is the string length and L is the sequence length. The probability that one position receives is regarded as the *coverage probability* of the position, which indicates the relative significance of the position for subtyping purposes. For each position in the multiple alignment of the 42 reference sequences by ClustalW (which was constructed through the EMBL-EBI server at 'http://www.ebi.ac.uk/clustalw/', in 113 min[1]), we computed its coverage probability and plotted them in Figure 5. One can see from this

[1]Unfortunately, we would not be able to construct the multiple alignment of all the 867 sequences due to the limit of sequence number in the server.We have also downloaded ClustalW for local testing but the estimated running time for the 867 sequences was more than two months on a desktop of 2.8GHz and 8GB memory. Nevertheless, we did submit a subset of 412non-C strains to the server, and obtained an MSA after more than 80 h.

plot that the most frequently covered positions by these 500 selected strings match very well with the most important positions in the ClustalW's multiple alignment (shown as red +, others as blue circles). This is another indication that our method is able to capture the critical sequential evolutionary information indicated by multiple sequence alignments.

In addition to the above string composition values computed within the second order Markov model, we have also examined the first order Markov model in which the composition value directly uses the string occurrence frequency. Also, in addition to the relative entropy scoring schemes, we have tested two other scoring functions: the SD of the composition values and the mean-weighted variant. It turned out that the relative entropy scoring scheme performed significantly better than the other two (data not shown); and the first order Markov model appeared much inferior to the second order Markov model (data not shown). Note that our string selection was done by examining all the strings within the length range, but assuming no correlations amongst them. One immediate future task is to consider possible correlations, and if identified, to borrow ideas from general feature selection methods and classification methods to exploit the feature correlations. Finally, but not of least interest, we will be working with other groups to further investigate the biological content of these 500 selected strings for post-subtyping studies. Interestingly, a BLAST search seemingly shows that the top ranked string GAAAAAGAG, by relative entropy, seems a signature of subtypes B, F and K (GAAAAAGAG appears in 41.8, 80, 100% subtype B, F, K strains in our dataset, respectively).

5 RECOMBINANT FORM PREDICTION

There are many proposed methods and programs which address RNA recombination detection, especially in RNA viruses, such as HIV and *hepatitis C virus* (HCV) (Martin *et al*., 2005; Milne *et al*., 2004; Rozanov *et al*., 2004). Most of these methods and programs start with multiple sequence alignments (Martin *et al*., 2005; Milne *et al*., 2004). For an HIV-1 viral strain known to be pure subtype, its subtyping is considered relatively easy. This has been demonstrated by our method, as well as the three subtyping programs we have tested. However, HIV-1 is notorious for its various forms of recombinations, which constantly challenge the drug development (Rambaut *et al*., 2004). Thus, upon the arrival of each new strain, the subtyping task is to determine whether it is a pure subtype strain, and, if it is not, to determine its recombinant form.

We randomly selected 16 out of the 825 pure subtype sequences, and for each of them, we partitioned it into 50 equal parts, each containing around 180 nt. At each testing, a consecutive ℓ parts, where $1 \leq \ell \leq 25$, were removed from the whole strain and the remainder were concatenated into a new sequence. Using the 5000 top ranked strings selected using the 42 HIV-1 pure subtype reference sequences, the new sequence was again represented as a 5000-dimensional composition vector. The distances between this vector and the 42 reference sequences were then calculated, and the subtypes of the two closest reference sequences were reported. For each strain, a total of 950 testings were executed and 1900 predicted subtypes were reported. For each of the distinct 13 subtypes, its occurrence frequency in these 1900 predicted subtypes was calculated, and every strain was represented as a 13D vector. These frequencies are color coded and plotted in Figure 6 (the left 16 columns), where one can easily see that those non-B strains are confidently evaluated to be pure subtype strains, though a few of them (two A1, two B and two D) have been mixed with some small percentage of information from other subtypes.

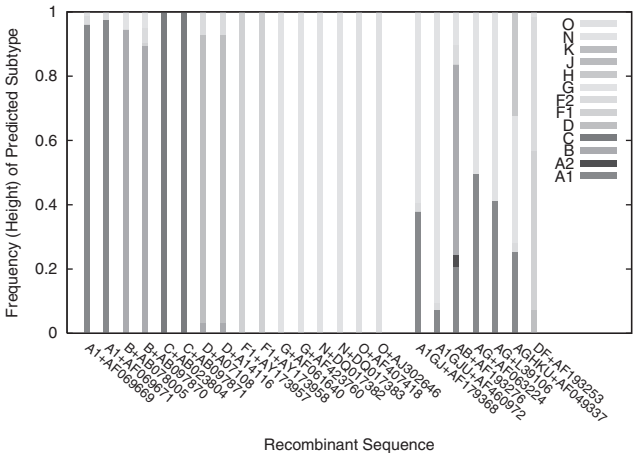

Fig. 6. The frequencies of the predicted subtypes for the 16 of 825 randomly selected pure subtype strains and the 7 recombinants, using the 5000 top ranked strings. Colour version of this figure is available as Supplementary material online.

We then used the above approach, extended from our pure subtyping method, to determine the HIV-1 circulating recombinant forms (CRFs). That is, each of the 331 HIV-1 recombinant strains (196 CRF01AE, 52 CRF02AG, 3 CRF03AB, 3 CRF04CPX, 3 CRF05DF, 8 CRF06CPX, 7 CRF07BC, 4 CRF08BC, 5 CRF09CPX, 3 CRF10CD, 10 CRF11CPX, 10 CRF12BF, 6 CRF13CPX, 7 CRF14BG, 5 CRF1501B (AE/B), 2 CRF16A2D, 4 CRF18CPX and 3 CRF19CPX) was partitioned into 50 equal parts, and there were 950 associated testings, for each of which the subtypes of the two closest reference sequences were reported. Then, similarly, for each of the distinct 13 subtypes, its occurrence frequency in these 1900 predicted subtypes was calculated and every recombinant strain was represented as a 13D vector (for seven randomly picked recombinant strains, their associated vectors are plotted in Figure 6, the right seven columns). The non-trivial portions of the predicted subtypes can be assigned as the recombinant form. For instance, for strain AF179368, 37.6% predicted subtypes are A1, 0.2% are F1, 2.6% are F2 and 59.6% are G. Therefore, we may predict this strain as A1G recombinant. The known recombinant form, recorded in the LANL HIV Sequence Database, is CRF11CPX and it is a mosaic of A/G/E/J. In other words, our computational prediction somehow missed subtype J information. For each recombinant strain, we calculated the prediction accuracy as the percentage of correctly assigned subtypes among the 1900 ones. The average prediction accuracy on those 91 recombinant strains (CRF02AG, CRF03AB, CRF05DF, CRF07BC, CRF08BC, CRF10CD, CRF12BF, CRF14BG and

CRF16A2D) which have the deterministic recombinant forms is 87.3%. Among the above seven selected recombinant strains, our method perfectly predicted on AF063224 and L39106, which are AG recombinants (Fig. 6, the 20th and 21st columns).

It is claimed that the NCBI genotyping tool is 'especially useful for the analysis of recombinant sequences' (Rozanov *et al*., 2004). To consider this claim, we first conducted a comparative study by submitting all the 331 recombinant strains to the server, but only allowed it to predict pure subtypes. For each sliding window, we reported the top two subtypes according to the BLAST similarity score, and similarly calculated the prediction accuracy. For the 91 strains having deterministic recombinant forms, the average prediction accuracy was 73.4%. Second, we replaced the 48 reference pure subtype strains in the tool by our 42 reference strains (in fact, our 42 are included in the 48) to test the BLAST methodology using the same set of reference strains. For the 91 strains having deterministic recombinant forms, the average prediction accuracy decreased a little to 66.2%. In another study, we allowed the server to report the closest recombinant form since it has reference recombinant strains. Among the collected 331 recombinant strains, 65 of them are used as references in the tool (the tool has in total 68 reference recombinant strains, in which 3 of them are absent from the LANL HIV Sequence Database). For the other 266 recombinant strains, the server made only two mistakes where two CRF12BF strains, AY771588 and AY771589, were predicted to pure subtype B. The prediction results remained exactly the same even when the 48 reference pure subtype strains were replaced by our 42 reference strains. In the last study to test our pure subtyping method, we used our 42 reference pure subtype strains and the 68 reference recombinant strains in the NCBI tool, and the 5000 selected nucleotide strings to map every strain into a 5000-dimensional space and subsequently calculated all the pair-wise distances. Each of the 266 testing recombinant strains was assigned the closest pure subtype or recombinant form. We were able to assign only 242 strains correctly, while all the other 24 strains were incorrectly assigned as CRF02AG. This indicates that the 5000 nt strings are not good enough for recombinant form prediction, since they were selected for pure subtyping purpose. Nevertheless, it is interesting to see that both AY771588 and AY771589 were predicted to CRF02AG but not pure subtype B, suggesting that the 5000 selected nucleotide strings might capture some information missed by BLAST.

6 CONCLUSIONS

We proposed a method to select the most informative strings and use only their composition values to represent the whole genomes. Such a proposal appears novel in the context of HIV-1 subtyping and recombinant form determination. It reduces the genomic data dimensionality, and possibly reduces sequential evolutionary noise, and thus makes feasible the whole genome phylogenetic analysis on a large set of sequences. Such a method also enables us to identify informative explicit strings with respect to a large set of sequences and therefore supports biological explanation. Using our method to select 500 strings, for a total 867 pure subtype HIV-1 viral strains, we were able to predict their subtype perfectly, i.e. 100% accuracy.

ACKNOWLEDGEMENTS

The authors are grateful to the research support from CFI, iCORE and NSERC. They also would like to thank two referees for their helpful comments.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.

Chen,X. *et al*. (2000) A compression algorithm for DNA sequences and its applications in genome comparison. In *Proceedings of the Sixth Annual International Computing and Combinatorics Conference (RECOMB)*. ACM Press. Tokyo, Japan. pp. 107–117.

deOliveira,T. *et al*. (2005) An automated genotyping system for analysis of HIV-1 and other microbial sequences. *Bioinformatics*, **21**, 3797–3800.

Dopazo,H. *et al*. (2004) Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. *Bioinformatics*, **20**, i116–i121.

Gifford,R. *et al*. (2006) Assessment of automated genotyping protocols as tools for surveillance of HIV-1 genetic diversity. *AIDS*, **20**, 1521–1529.

Grumbach,S. and Tahi,F. (1994) A new challenge for compression algorithms: genetic sequences. *J. Inf. Proces. Manage*., **30**, 875–866.

Hao,B. and Qi,J. (2003) Prokaryote phylogeny without sequence alignment: from avoidance signature to composition distance. In *Proceedings of the 2003 IEEE Bioinformatics Conference (CSB 2003)*, pp. 375–385.

Herniou,E. *et al*. (2001) Use of whole genome sequence data to infer baculovirus phylogeny. *J. Virol*., **75**, 8117–8126.

House,C. and Fitz-Gibbon,S. (2002) Using homolog groups to create a whole-genomic tree of free-living organisms: an update. *Mol. Evol*., **54**, 539–547.

Karlin,S. and Burge,C. (1995) Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*., **11**, 283–290.

Leitner,T. *et al*. (2005) *HIV-1 Subtype and Circulating Recombinant Form (CRF) Reference Sequences*. Accessible through http://www.hiv.lanl.gov/content/hiv-db/REVIEWS/RefSeqs2005/RefSeqs05.html.

Li,W. *et al*. (2002) Phylogeny based on whole genome as inferred from complete information set analysis. *J. Biol. Phy*., **28**, 439–447.

Martin,D.P. *et al*. (2005) RDP2: recombination detection and analysis from sequence alignments. *Bioinformatics*, **21**, 260–262.

Milne,I. *et al*. (2004) TOPLi: software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics*, **20**, 1806–1807.

Milosavljevic, A. (1993) Discovering sequence similarity by the algorithmic significance. In *Proceedings of the 1st International Conference on Intelligent Systems for Molecular Biology (ISMB)*, pp. 284–291.

Myers,R.E. *et al*. (2005) A statistical model for HIV-1 sequence classification using the subtype analyser (STAR) *Bioinformatics*, **21**, 3535–3540.

Qi,J. *et al*. (2004) Whole proteome prokaryote phylogeny without sequence alignment: a *k*-string composition approach. *J. Mol. Evol*., **58**, 1–11.

Rambaut,A. *et al*. (2004) The causes and consequences of HIV evolution. *Nat. Rev. Gene*., **5**, 52–61.

Rivals,E. *et al*. (1996) Compression and genetic sequences analysis. *Biochimie*, **78**, 315–322.

Rozanov,M. *et al*. (2004) A web-based genotyping resource for viral sequences. *Nucleic Acids Res*., **32**, W654–W659.

Saitou,N. and Nei,M. (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. and Evol*., **4**, 406–425.

Snel,B. *et al*. (1999) Genome phylogeny based on gene content. *Nat. Genet*., **21**, 108–110.

Snel,B. *et al*. (2000) Genome evolution: gene fusion versus gene fission. *Trends Genet*., **16**, 9–11.

Stuart,G. and Berry,M. (2003) A comprehensive whole genome bacterial phylogeny using correlated peptide motifs defined in a high dimensional vector space. *J. Bioinform. and Comput. Biol*., **1**, 475–493.

Stuart,G. *et al*. (2002a) A comprehensive vertebrate phylogeny using vector representation of protein sequences from whole genomes. *Mol. Biol. Evol*., **19**, 554–562.

Stuart,G. *et al*. (2002b) Integrated gene and species phylogenies from unaligned whole genome sequence. *Bioinformatics*, **18**, 100–108.

Stuart,G. *et al*. (2004) A whole genome perspective on the phylogeny of the plant virus family *tombusviridae*. *Arch. Viro*., **149**, 1595–1610.

Su,A.I. *et al*. (2001) Molecular classification of human carcinomas by use of gene expression signatures. *Cancer Res*., **61**, 7388–7393.

Wu,X. *et al*. (2006) Whole genome phylogeny construction via complete composition vectors. *Int. J. Bioinform. Res. Appl*., **2**, 219–248.