A time lag insensitive approach for estimating HIV-1 transmission direction

Jing Yang[a], Meng Ge[b] and Xian-Ming Pan[a]

Objectives: Identifying the direction of transmission in transmission pairs is important both for forensic investigations and for the monitoring of HIV epidemics, however, reliable methods are not yet available due to the long time lag between infection and sampling in most real cases.

Designs: Based on bottleneck effect and coreceptor switching, we aimed at identifying an estimator from sequences of viral gp120 proteins to determine transmission direction between transmission pairs. The estimator should be changed with HIV transmission but was independent of disease progression in an individual.

Methods: Here, we present a novel and reliable approach for identifying transmission direction. We derived a set of conserved patterns, called common patterns, from the sequences of viruses, which differed in their coreceptor usage. The number of unique common patterns in viral sequences decreased with transmission but remained almost constant with the progress of disease in an individual. We used this number as an estimator to determine transmission direction in 73 transmission pairs for which the transmission direction was already known.

Results: Our method predicted transmission direction with an accuracy of up to 94.5%. Of greater importance, our approach was not influenced by time lags between infection and sampling, and even transmission direction for transmission pairs with long time lags ranging from 2 years to more than 18 years were correctly determined.

Conclusion: Our approach for accurately determining transmission direction between transmission pairs is irrespective of the time lag between infection and sampling, which means a promising applications prospect.

AIDS 2012, **26**:921–928

Keywords: criminal cases, HIV epidemics, HIV transmission direction, patterns, sampling time lags, transmission pairs

Introduction

Since the first highly published case of the intentional transmission of HIV viruses in the 1990s, in which six patients became HIV-1 positive after being treated by a dentist who was knowingly HIV-1 positive [1–3], such cases have been found in many countries around the world. In 2007, United Nations Programme on HIV/AIDS and United Nations Development Programme raised concerns about decisions reached in these criminal cases [4], highlighting the importance of molecular evidence of transmission direction to strengthen judgments made on the identification of transmission sources. The use of molecular evidence for identifying transmission direction is also important for identifying the characteristics of HIV transmission networks, which affect the rate of disease transmission in the short term and the prevalence of the disease in the long term [5]. Currently, phylogenetic analysis of HIV-1 sequences, based on assessing the similarity of viral sequences in transmission partners, is used widely to determine HIV transmission linkage, but reliable

[a]Key Laboratory of Bioinformatics, Ministry of Education, School of Life Sciences, Tsinghua University, and [b]National Laboratory of Biomacromolecules, Center for Structure and Molecular Biology, Institute of Biophysics, Chinese Academy of Sciences, Beijing, China.

Correspondence to Xian-Ming Pan, School of Life Sciences, Tsinghua University, Beijing 100084, China.
Tel: +86 10 627 928 27; fax: +86 10 627 928 27; e-mail: pan-xm@mail.tsinghua.edu.cn
Received: 15 October 2011; revised: 3 February 2012; accepted: 7 March 2012.

DOI:10.1097/QAD.0b013e3283536b89

identification of transmission direction is still not possible [6,7].

Recently, some attempts have been made to identify transmission direction using phylogenetic analysis of paraphyletic relationships [8]. During viral transmission, only a few viral isolates are transmitted from source to recipient, the so-called bottleneck effect [9]. As a result, only a subset of source sequences will be more closely related to all recipient sequences than all source sequences to each other. This relationship between source and recipient sequences is thus termed a paraphyletic relationship, and provides molecular evidence for demonstrating transmission direction. However, due to the rapid evolution of HIV viruses, paraphyletic relationships are gradually lost through time. This is a particular problem because, in most cases, there are long time lags between transmission and sampling [7]. For example, in almost all criminal cases, the length of time between transmission and DNA testing of the suspect is extensive due to delays in reporting and detecting crime. Therefore, phylogenetic analysis of paraphyletic relationships may not be effective or reliable in most real situations, and the development of alternative methods is necessary.

It is well known that viral coreceptor usage switching is a common step in the progression of AIDS. In the early stages of infection, viruses select CCR5 as their coreceptor for entering the host cell, but subsequently switch to coreceptor CXCR4. Generally speaking, there is an intermediate period when viruses can bind to either CCR5 or CXCR4 to facilitate their entry into host cells [10]. By analysing coreceptor usage switching, it may be possible to identify markers that are affected by transmission but remain constant throughout the progress of disease in an individual. Such disease progression-independent markers could be used to develop methods for identifying transmission directions that are not affected by time lags in sampling.

In this work, we developed an approach for identifying transmission direction based on a subset of 'common patterns' in the HIV-1 gp120 protein derived from CCR5/CXCR4 coreceptor usage-labeled sequence datasets. To verify our approach we identified the transmission directions of 73 transmission pairs for which the transmission direction was already known. Our approach identified transmission direction with an accuracy of up to 94.5%. It performed even better on transmission pairs with longer sampling time lags, and was not influenced by viral subtype or transmission route.

Methods

Derivation of common patterns

We searched the Los Alamos HIV-1 databases (http://www.hiv.lanl.gov/; last modified 26 January 2011), and collected all sequences in the *env* C2-V5 region with lengths of about 180 amino acids (genomic region 7050-7590), which were labeled coreceptor usage. The following dataset was constructed.

Dataset 1: coreceptor usage dataset

In total, there were 1926 sequences from 528 patients (on average, 3.6 sequences per patient), of which 1485 sequences of viruses only used CCR5 as the coreceptor (termed R5 sequences) and 441 sequences of viruses used other coreceptors (termed non-R5 sequences). The non-R5 sequences included 167 sequences of viruses, which used only CXCR4 as the coreceptor (termed X4 sequences) and 274 sequences of viruses, which used either CCR5 or CXCR4 as the coreceptor (termed R5X4 sequences).

Definition of patterns

We defined a 'pattern' as a group of nonsequential but related amino acids; for a given subsequence window of length L, a pattern is a sequence of m residues, the first residue of which is fixed in the first position at the left-hand side of the window, and the remaining m-1 residues are distributed in the remaining positions of the window. The number of possible combinations of positions for the m residues (denoted as s) in the subsequence is:

$$s = \frac{\prod_{k=1}^{m-1}(L-k)}{(m-1)!} \quad (1)$$

Here, we set the number of letters per pattern (m) to 4, and the length of the subsequence window (L) to 20 to search for patterns in the viral sequences.

There are 20 possible amino acid letters for each position, giving a total of $s \times 20^m$ possible patterns. The subsequence window was used to sequentially search along each sequence step by step, to obtain all patterns.

Common patterns and most recent common ancestor patterns

'Common patterns' are those that appear in both R5 sequences and non-R5 (R5X4 or X4) sequences. To avoid the random generation of erroneous patterns, only those patterns appearing in more than 60 R5 sequences and at least one non-R5 sequence were defined as common patterns.

For comparison, we also defined another subset of patterns appearing in R5 sequences, but not in R5X4 and X4 sequence, called 'most recent common ancestor patterns' (MRCA patterns). To avoid the random generation of erroneous patterns, only those MRCA patterns, which appeared in more than 60 R5 sequences, were chosen.

Conservation of common patterns during disease progression in an individual

In order to test the stability of the number of unique common patterns during disease progression in an individual, a set of sequences derived from samples taken from longitudinally observed patients was constructed as described below.

Dataset 2: longitudinal sampling dataset

A set of sequences from the *env* C2-V5 region (amino acids approximately 260–470) from nine serially sampled infected patients (currently available patients who were sampled at more than 10 time points) whose progression to AIDS occurred in the same year were obtained from GenBank (Accession numbers AF137629 to AF138163, AF138166 to AF138263, and AF138305 to AF138703). Samples were obtained at roughly 6 monthly intervals ranging from 0 to 11 years postseroconversion [11]. Samples with less than five viral sequences were excluded. The numbers of unique MRCA patterns and unique common patterns appearing in viral sequence sets at different sampling time points were calculated for each of these nine patients.

Transmission direction identification

As a consequence of the bottleneck effect, only a subset of common patterns is transmitted from source to recipient. Therefore, the number of unique common patterns tends to decrease during transmission from source to recipient. The number of such patterns in a viral sequence set should, therefore, be a suitable estimator of transmission direction.

In order to test whether the number of unique common patterns is a suitable estimator of transmission direction or not, 73 transmission pairs (each pair containing a source and a recipient) of known transmission direction were identified, and their sequences were collected. We searched the Los Alamos HIV-1 databases and collected all sequences in the *env* C2-V5 region with lengths of about 180 amino acids (genomic region 7050–7590) whose cluster transmission type was labeled as 'Mother→Child', 'Heterosexual', or 'Men sex with men'. Sequences collected for each patient were assembled according to time point into viral sequence sets, based on the original papers, sequence filenames or comments in database files. Samples with less than five viral sequences were excluded. We then referred back to the original papers to confirm the transmission direction of all partners [12–26]. A total of 73 pairs had clear transmission linkages and direction, of which 53 had mother-to-child transmission, 14 were heterosexual partners and six were homosexual partners. Detailed information on these 73 transmission pairs is shown in Supplemental Digital Content 1, http://links.lww.com/QAD/A214. In most cases, data for each patient fell into more than one viral sequence set. In order to comprehensively test the performance of our approach, we used two entirely different methods to select a single time point for each patient, yielding the following two datasets.

Dataset 3: 'minimal time lag dataset'

In this dataset, time points selected for each patient were those that minimized the difference in sampling time between the sample derived from the source and the sample derived from the recipient, and were closest to the transmission time point.

Dataset 4: 'maximal time lag dataset'

In this dataset, time points selected for each patient were those that maximized the difference in sampling time between the sample derived from the source and the sample derived from the recipient, and were furthest from the transmission time point.

The number of unique common patterns appearing in a viral sequence set was used as the estimator of transmission direction. Transmission direction was determined respectively for the above two datasets. It should be noted that when a pattern appeared in a viral sequence set, it was only counted once, irrespective of whether it appeared only once or many times.

Weight score and 10-fold cross validation

In order to improve the performance of our approach, each pattern was given a weighted score, W. The weighted score of the mth pattern was defined as the following:

$$W_m = \sum \frac{n_0 - n_1}{N} \tag{2}$$

Wherein, n_0 is the total number of sources containing the mth pattern, and n_1 is that of recipients. N is the total number of transmission pairs. Thus, if the pattern appears more times in viral sequence sets from sources than in viral sequence sets from recipients, its weighted score will be positive; otherwise its weighted score will be negative.

The sum of weighted scores for patterns appearing in a given viral sequence set, SC_k, should be a better estimator for identifying transmission direction.

$$SC_k = \sum_{m=1}^{ss} W_m \bullet E_{mk} \tag{3}$$

Wherein, SC_k is the score of kth viral sequence set; E_{mk} has values of either 1 or 0 and represents whether or not the mth pattern appears in the kth viral sequence set, and ss is the total number of patterns.

In order to test the applicability of our approach to other datasets, 10-fold cross validation was performed. All 73 transmission pairs were randomly partitioned into 10