# Analysis of Within-host Viral Evolution by Reconstructing the Longitudinal Phylogenetic Tree of HIV-1

Hiroshi Tanaka
*Department of Bioinformatics, Tokyo Medical and Dental University*
*1-5-45 Yushima, Bunkyo-ku, Tokyo, JAPAN*
*tanaka@cim.tmd.ac.jp*

Fengrong Ren and Soichi Ogishima
*Department of Bioinformatics, Tokyo Medical and Dental University*
*1-5-45 Yushima, Bunkyo-ku, Tokyo, JAPAN*
*{ren, ogishima}@bioinfo.tmd.ac.jp*

## Introduction

Many of worldwide genomic projects are ongoing with very rapid speed, which expand our genomic knowledge on various medical problems. With this background, we have been involved in analyzing the longitudinal process of within-host molecular evolution of various viruses. In this study, we propose a new method for the analysis of the within-host viral evolution. This approach is based on the sequential viral samples obtained at different time points from the same host and reconstructs a longitudinal phylogenetic tree that describes the bifurcation, diversification, and extinction of viral evolution. Furthermore, since the evolutional model employed in this approach is so flexible to apply to all the modes of molecular evolution such as purifying, neutral and adaptive evolution, it can estimate when and in which mode the virus evolves. A data set of HIV-1 envelope gene is applied to examine the efficiency of this method.

## Data and Method

### Data

The sequences analyzed in this study are from GenBank, which were sequentially isolated over 7 years after HIV-1 infection of one single patient [1]. In the original paper, there were a total of 24 different V3 loop sequences (105bp) and they were assigned letters A-F. In year 0 (1984), only one sequence was observed and assigned letter A. No data were available for years 1 and 2 (1985 and 1986). The other 23 sequences were obtained from year 3 to year 7 (1987, 1988, 1989, 1990 and 1991), and they are sequences B, C1-5, D1-8, E1-8 and F. This patient remained asymptomatic and had not received the antiviral therapy during the period when the samples were obtained

### Condon-based Model

For analysis of molecular evolution, the traditional methods are developed according to the neutral evolutionary theory, and have difficulty in detecting positive selection. Our method developed in this study can resolve this problem. Since the excess of nonsynonymous substitutions over synonymous substitutions is an indicator of positive selection, the estimation of nonsynonymous ($d_N$) and synonymous ($d_s$) substitution rates is very important. For calculating $d_N$ and $d_s$, we employed the codon-based model of Goldman and Yang [2-4]. This model is formulated at the codon level instead of the nucleotide or amino acid level and is thus more realistic compared to other evolutionary models.

### Distance-based linking Algorithm

To follow-up the changes of viral evolution in molecular level, we need to analyze the sequential viral samples. The sequential samples are the sequences that are isolated from a single patient at different time points and usually represent the changes of phylogenetic relations between viral variants along with the passage of time. However, the traditional methods developed for reconstructing molecular phylogenetic trees, such as maximum likelihood (ML) and neighbor-joining (NJ), cannot deal with this kind of data, because they were designed for sequence data obtained at the same time point. Therefore, we developed a new algorithm to link variants observed at different time points [5][6]. The details of this algorithm are as follows:

**Step 1**: calculate the distances between all variants observed at the time point $T_i$ and $T_{i+1}$ using NJ method.

**Step 2**: chose the $T_i$-variants whose nearest variant is a $T_{i+1}$-variant rather than a $T_i$-variant as the immunological escape variants.

**Step 3**: reconstruct the phylogenetic tree using the immunological escape $T_i$-variants and $T_{i+1}$-variants.

**Step 4** cut the longest branch and divide the variants of $T_i$ and $T_{i+1}$ into different groups. If the number of immunological escapes $T_i$-variants > 2, we should cut the longer branches more than one time. Namely, if the number of immunological escape is $n$, the number of longer branches we should cut is $n$-1.

**Step 5**: reconstruct the phylogenetic trees of each group using NJ method, and then the relations between $T_i$ and $T_{i+1}$ variants would be clear.

## Results

The tree reconstructed by our method is shown in figure 1. The tree topology is very similar to that of Holmes et al. (1992) obtained by calculating the differences between the amino acid sequences by hand. As no data were available for years 1 (1985) and 2 (1986), new viral variants appeared from year 3 (1987). They were mainly divided into groups C (including C1, C2, C3, C4 and C5) and E (only sequences E1 was observed in year 3), and sequence F. In year 4 (1988), the five sequences of group C disappeared due to extinction or too lower frequencies, and the new group D (including D1, D2 and D3) diverged from sequences C3. E1 was not observed as well, but similar sequences E2 and .E3 were observed. In year 5 (1989), instead of D1, D2 and D3, the new variants D4 and D5 were observed within group D. On the other hand, E1 (revival?), E2 and new variants E4 and E5 are observed within group E. In year 6 (1990), the new variants D6, D7 and D8 were observed instead of D4 and D5, and E6, E7 and E8 were observed instead of E2, E4 and E5. However, in year 7 (1991), only sequences D7, D8 and E1 were observed, and besides these three sequences, sequence F, which disappeared for several years, was observed again in this year. We also used traditional methods; maximum likelihood and neighbor-joining, to calculate this data set, but the results are quite different from that of Holmes et al. (the results are not shown here). Both of them could not reconstruct the phylogenetic tree, which correctly show the evolutionary relations between these 24 viral sequences.

Furthermore, since we incorporate a codon-based model in our method, not only is the tree topology constructed, but also the evolutionary patterns in different phases of viral evolution process are detected.

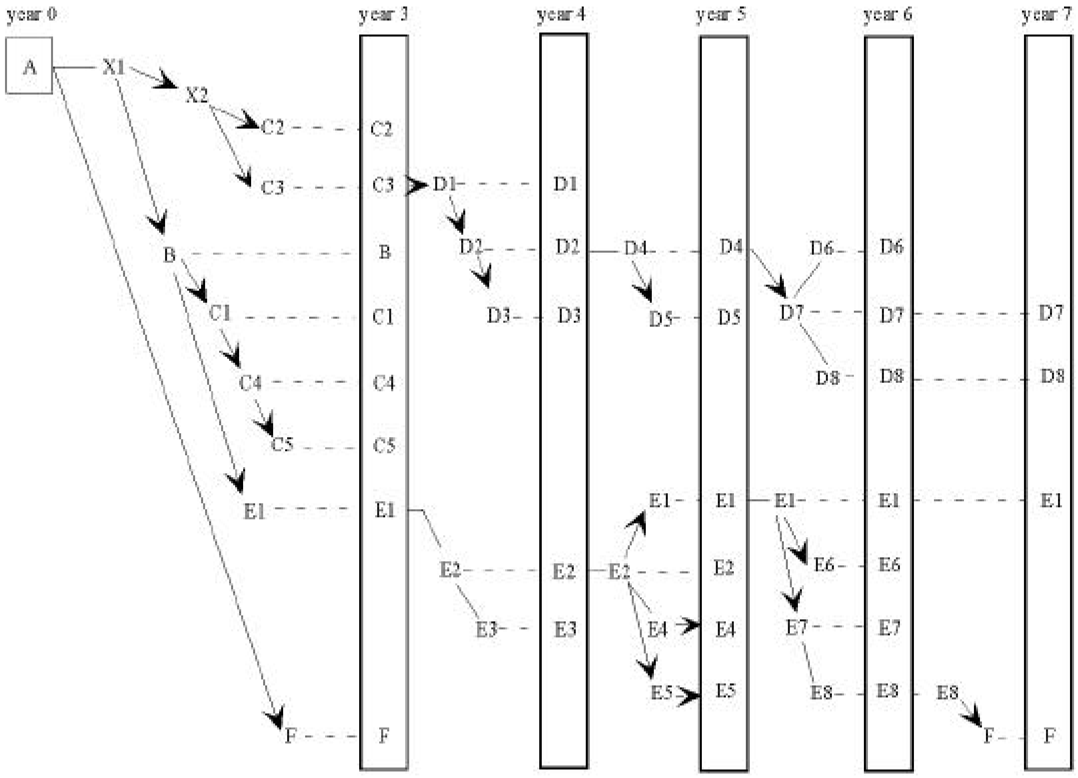

**Figure 1**: Longitudinal phylogenetic tree of 24 sequential sequences of HIV-1 *env* gene reconstructed by our method. Solid lines without arrow represent purifying selection, whereas the solid lines with arrow represent positive selection. Dotted lines represent identical continuation of viral variants.

## Discussion

We developed a new approach to infer the sequential viral samples and incorporated a "codon-based model" into this approach. We applied this approach to a data set of HIV-1 *env* gene and reconstructed a longitudinal phylogenetic tree that can describe the evolutionary process of within-host virus. Using this approach, not only the phylogenetic relations between viral variants but also the patterns of within-host evolution can be estimated. The results obtained in this study show that the purifying selection and adaptive evolution appear to occur alternately in the process of viral evolution. In the beginning of infection (from 1984 to 1986), nonsynonymous substitutions were hardly observed in sequence A and the viral sequence evolved in neutral form. However, in year 3 (1987), the nonsynonymous substitutions rapidly increased and exceeded synonymous substitutions, with $d_N/d_S$ ratio well above one. These results strongly suggest that positive selection operated in this period. As a consequence, the viral sequence evolved into two groups, C and E, and a number of diverse sequences including quasi-species sequences (within one group) appeared. Obviously, year 3 is a period of positive selection. Positive selection is also observed in years 4 (1988) and 5 (1989); as a result, a new group, D, appeared. From year 6 (1990), however, the nonsynonymous substitution rate was rapidly reduced and the viral sequences switched to purifying selection again. We think that such kind of dynamic process, that is purifying selection (or neutral evolution) - adaptive evolution - purifying selection, probably described the evolutionary process of within-host virus realistically.

The results obtained in this study suggest that our new method may provide a more realistic description of viral evolutionary process than traditional methods. First, this method can distinguish purifying selection and positive selection, whereas traditional methods assume that only neutral evolution occurs. Second, the sequential-linking approach developed in this study makes it possible to deal with sequential viral sequences determined at different time points and to reconstruct a longitudinal phylogenetic tree that can more accurately describe the phylogenetic relationships of sequential data. Our method exploits information contained in temporal observation within a host more thoroughly.

## Acknowledgments

This study is supported by Grand-in-Aid for Scientific Research from Ministry of Education, Science, Sports and Culture of Japan.

## REFERENCE

[1] E. C. Holmes et al, "Convergent and divergent sequence evolution in the surface envelope glycoprotein of human immunodeficiency virus type 1 within a single infected patient" *Proc. Natl. Acad. Sci. USA* 1992: 89 pp4835-4839

[2] N. Goldman and Z. Yang, "A codon-based model of nucleotide substitution for protein-coding DNA sequence" *Mol. Biol. Evol.* 1994: 11 pp725-736

[3] Z. Yang and R. Nielsen, "Synonymous and nonsynonymous rate variation in nuclear genes of mammals" *J. Mol. Evol.* 1998: 46 pp409-418

[4] Z. Yang, "PAML: A program for package for phylogenetic analysis by maximum likelihood" *CABIOS* 1997: 15 pp555-556

[5] F. Ren, H. Tanaka and T. Gojobori: Construction of molecular evolutionary phylogenetic trees from DNA sequences based on minimum complexity principle, *Computer Methods and Programs in Biomedicine*, 1995: 46 pp121-130

[6] H. Tanaka, F. Ren, T. Okayama and T. Gojobori, Topology Selection in Unrooted Molecular Phylogenetic Tree by Minimum Model-Based Complexity Method, *Biocomputing'99*, World Scientific 1999: pp326-337

## RESUME

Un nouvel algorithme pour inférer « l'évolution de dans – hôte » des séquence virale est présenté. L'approche de « séquentiel – lier » est développé afin que l'arbre phylogénétique longitudinal peut être reconstruit à partir de données de séquences moléculaires qui est obtenu à points du temps différents du saml'hôte. L'algorithme emploie un modèle de base - codons qui utilise un processus de Markov décrire des substitutions entre codons, à calculat,nonsynonymous et la substitution synonyme estime et distinguer le positivsélection et évolution neutre. L'algorithme est appliqué à une série de gènes de la région V3 de l'enveloppe du HIV - 1 séquencés à différentes années après l'infection d'un seul patient.