Pacific Symposium on Biocomputing 6:595-605 (2001)

A NEW ALGORITHM FOR ANALYSIS OF WITHIN-HOST HIV-1 EVOLUTION

F. REN, S. OGISHIMA, H. TANAKA
Department of Bioinformatics, Medical Research Institute
Tokyo Medical and Dental University
Yushima 1-5-45, Bunkyo, Tokyo 113-8510, Japan

A new algorithm for inferring the evolution of within-host viral sequences is presented. A sequential-linking approach is developed so that a longitudinal phylogenetic tree can be reconstructed from sequential molecular data that are obtained at different time points from the same host. The algorithm employs a codon-based model, which uses a Markov process to describe substitutions between codons, to calculate nonsynonymous and synonymous substitution rates and to distinguish positive selection and neutral evolution. The algorithm is applied to a data set of the V3 region of the HIV-1 envelope genes sequenced at different years after the infection of a single patient. The results suggest that this algorithm may provide a more realistic description of viral evolution than traditional evolutionary models, because it accounts for both neutral and adaptive evolution, and reconstructs a longitudinal phylogenetic tree that describes the dynamic process of viral evolution.

1 Introduction

Since Kimura's neutral theory [1] has been accepted by biologists for the mechanism of evolution at the molecular level, most methods for analyzing molecular evolution are developed based on this theory. In contrast to Darwin's theory of evolution by natural selection, the neutral theory maintains that molecular evolution is mainly caused by random fixation of neutral mutations and functional constraint rather than by positive selection fixing advantageous mutations. However, positive selection was also recently observed at the molecular level by many biologists, and a number of methods have been developed for its detection [2-5]. These studies, however, aim to detect positive selection, and do not study how neutral evolution and positive selection behave as time goes on. In this study, we develop a new method to detect when the neutral and adaptive evolution occur in the process of within-host viral evolution, and to reconstruct a longitudinal phylogenetic tree using sequential molecular sequences. We applied the method to a data set of the V3 region of HIV-1 envelope gene and carried out a detailed analysis of viral evolution.

There are two reasons why we applied our method to HIV data: one is that it is well known that the evolutionary rate of virus is very rapid compared to most DNA genomes. The evolutionary rate of retrovirus is estimated to be millionfold greater than that of host DNA genomes. This feature of virus is quite appropriate for analysis of real-time molecular evolutionary process. The other is that the evolutionary pattern the virus would take in different stages of infection is closely

related to the immunological status of host. Therefore longitudinal phylogenetic analysis is especially significant to medical assessment of the disease.

2 Data and Methods

2.1 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 [6]. 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.

2.2 Condon-based Evolutionary Model

Traditional methods for analysis of molecular evolution are developed according to the neutral evolutionary theory, and have difficulty in detecting positive selection. Our method developed in this paper can resolve this problem. The excess of nonsynonymous substitutions over synonymous substitutions is an indicator of positive selection, so that estimation of nonsynonymous (d_N) and synonymous (d_s) substitution rates, is very important. For calculating d_N and d_s, the codon-based model of Goldman and Yang [7] is employed in this study. The 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. In this model, substitutions between sense codons are described by a continuous-time Markov process. The three stop codons are not allowed in the gene sequence and are thus not considered in the Markov model. The Markov process is described by a rate matrix $Q = \{q_{ij}\}$, where $q_{ij}\Delta t$ represents the probability that codon i changes to codon j in a small time interval Δt. By using the rate matrix Q, we can describe the codon substitution matrix $P(t)$, whose element $p_{ij}(t)$ gives the probability that codon i replaces codon j after time interval t. According to Markov process theory, we have

$$P(t) = e^{Qt} .$$

The instantaneous substitution rate from codons i to j ($i \neq j$) is given by

$$q_{ij} = \begin{cases} 0, & \text{if } i \text{ and } j \text{ differ at two or three codon positions,} \\ \mu\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transversion,} \\ \mu\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a synonymous transition,} \\ \mu\omega\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transversion,} \\ \mu\omega\kappa\pi_j, & \text{if } i \text{ and } j \text{ differ by a nonsynonymous transition,} \end{cases}$$

where parameter κ is the transition/transversion rate ratio, ω is the nonsynonymous/synonymous rate ratio (= d_N/d_S), π_j is the equilibrium frequency of codon *j*, calculated using the nucleotide frequencies at the three codon positions, and μ is a scale factor, chosen such that time *t* is measured by the expected number of nucleotides substitutions per codon. Values of ω = 1, > 1, and < 1 mean neutral evolution, positive selection, and purifying selection on the protein, respectively. This model assumes that mutations occur at the three codon positions independently and only single-nucleotide changes are permitted instantaneously.

2.3 Sequential-linking Algorithm

Problems in phylogenetic analysis There are two problems that cannot be resolved by current methods in phylogenetic analysis of viral molecular sequences. One is how to estimate the phylogenetic relations between sequential sequences. These sequences are isolated from a single patient at different time points and usually represent the changes of phylogenetic relations between viral variants along

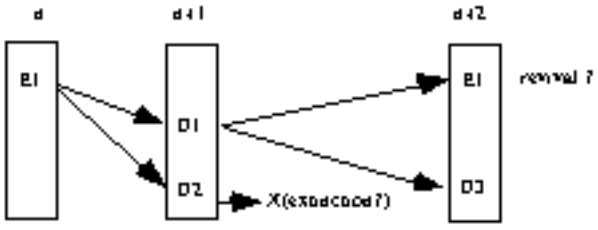

Fig. 1. *A model of changing phylogenetic relationships of viral variants over time points n, n + 1, … . E1, D1, D2 and D3 represent different variants.*

with the passage of time (see figure 1). However, traditional methods developed for reconstructing molecular phylogenetic trees, such as maximum likelihood (ML) [8] and neighbor-joining (NJ) [9], cannot deal with this kind of data, because they were designed for sequence data obtained at the same time point (see figure 2).

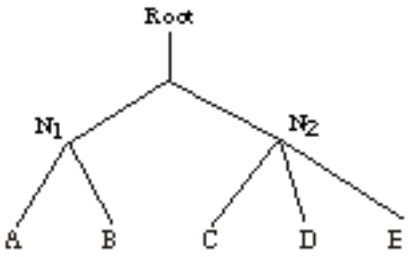

Fig. 2. *Phylogenetic tree reconstruction by traditional methods. A, B, C, D and E represent sequences of homologous genes from five species. Sequences at ancestral nodes N_1 and N_2 are unknown and can only be estimated from current species.*

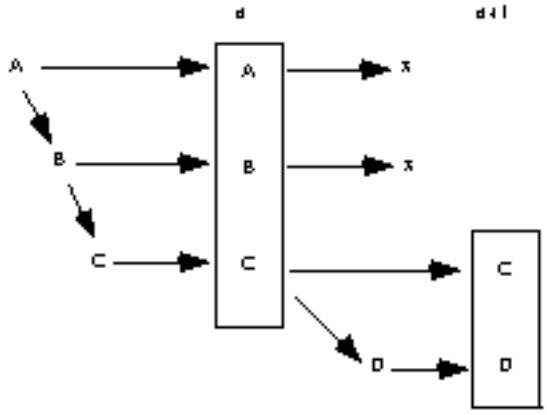

Fig. 3. *Branching time points, extinction time points, and observation time points of viral variants.*

Another problem is that the traditional methods overlook much of useful information that exists in sequential observations in estimating the phylogenetic relationship. For example, the real branching time of two variants may not be equal to the observed branching time point, because they might have branched between two observed time points. Likewise, some viral variants might become extinct and some might revive between the observed time points (revival or extinction might not really occur, and the absent of observation of the viral variant might be due to its very low frequency in blood). Viral variants might change (branch, go to extinction and revive) between the observed time points (see figure 3) and such

information should be used in studying viral evolution. Our new method makes use of such additional information.

Algorithm We developed a new algorithm to link variants observed at different time points. The details of this algorithm are as follows

(1) Construct phylogenetic trees T_n and T_{n+1} from the sequences of viral variants observed at time points *n* and *n+1*, separately. The CODEML program in PAML [7] is used to calculate synonymous and nonsynonymous substitution rates, and the tree topology is estimated by using ML method (DNAML program in PHYLIP)

(2) Choose the best link between variants observed at time points *n* and *n+1*, that is the tree having the highest likelihood among trees generated in the following steps (figure 4):

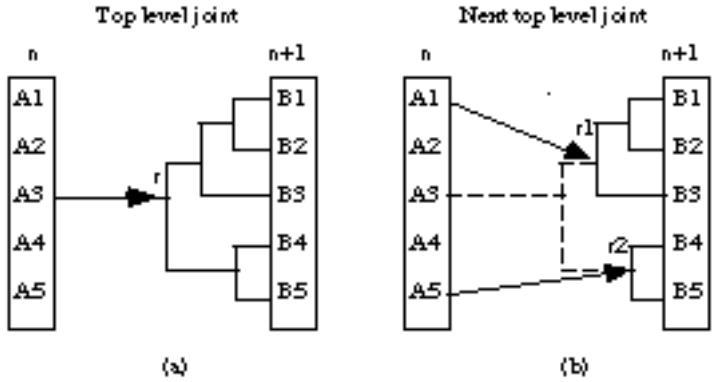

Fig. 4. *The sequential-linking approach developed in this study.*

- Step 1: choose one variant of T_n each time and link it with the top-level node of the tree T_{n+1} (the root of T_{n+1}) and calculate the maximum likelihood value of each linking tree topology thus generated. Take figure 4a an example, we link the top-level node of T_{n+1}, r, with A1, A2, A3, A4 and A5, respectively and five tree topologies are generated. Then, these tree topologies are used to calculate the likelihood and substitution rates (synonymous and nonsynonymous) by CODEML program, and the tree having the highest likelihood is selected.
- Step 2: decompose T_{n+1} into the next top-level and link each of its nodes to the variants of T_n in the same way as in step 1. In figure 4b, each node of the next top-level of T_{n+1}, r1 and r2, is linked with A1, A2, A3, A4 and A5,

respectively. Then the generated tree topologies are evaluated, and the best one among them is selected. In this step, all linking tree topologies including those generated in step 1 are compared.

- Step 3: do a further decomposition to T_{n+1} and repeat the procedure of link and calculation described in step 2, and compare all the linking tree topologies generated in each step. If the linking trees generated in step 3 do not show higher likelihood value than those generated in step 2, the decomposition procedure will be terminated.

(3) If the variants are observed at both time points *n* and *n+1*, we link them directly.

If some estimated branch lengths are extremely small [10], these branches can be considered as an internal node. For example, branches C1, C2 and C4 in figure 5a are extremely short and are collapsed in figure 5c.

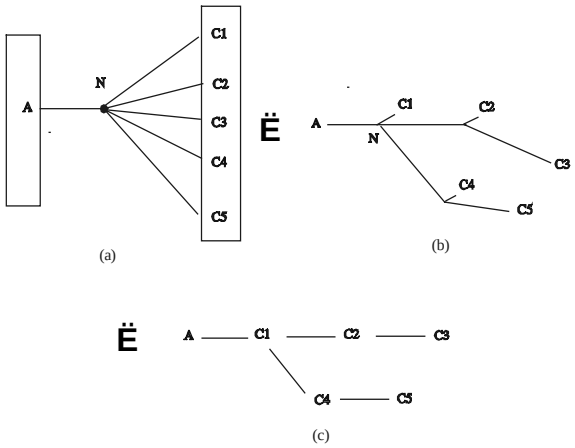

Fig. 5. *A phylogenetic tree model*

3 Results

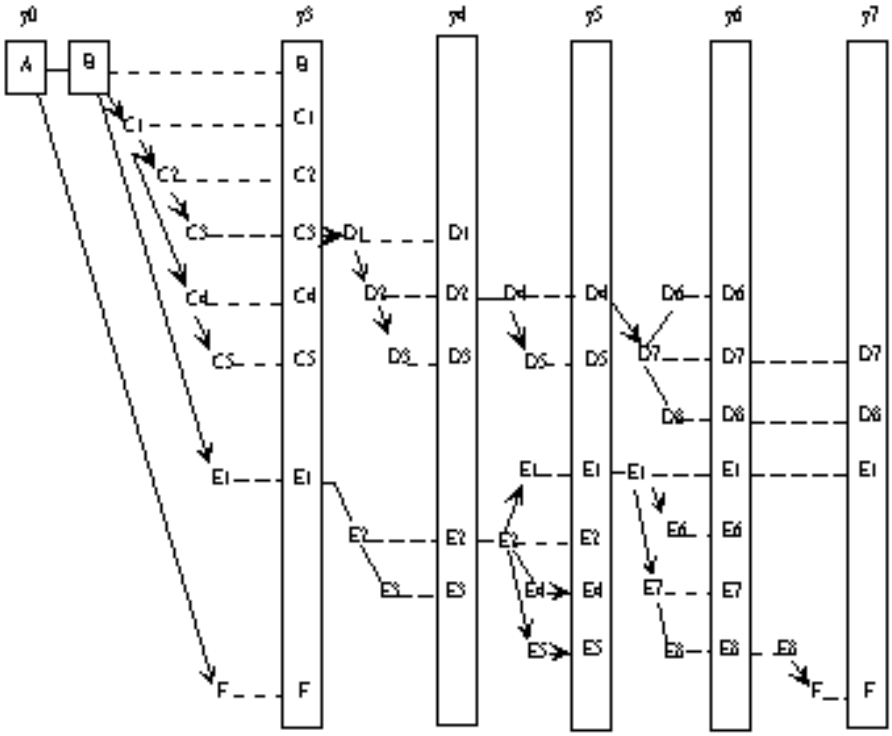

Fig.6. *Longitudinal phylogenetic tree of 24 sequential sequences of HIV-1 env gene reconstructed by our method. Solid lines without arrow represent purifying selection, whereas the solid lines with arrow represent positive selection. Dotted lines represent identical continuation of viral variants.*

The tree reconstructed by our sequential-linking method is shown in figure 6. 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 and 2, new viral variants appeared from year 3. 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, 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, 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, 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, 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.

However, the two phylogenetic trees reconstructed by traditional methods, maximum likelihood and neighbor-joining, show quite different results from that of Holmes et al. (see figure 7 and 8). The ML tree appears to have too many small branches so that the tree shows unnecessarily linked shape. In contrast, the NJ tree appears to have too many divergent branches so that the tree shows unnecessarily branched shape. Consequently, both of them could not estimate correct phylogenetic relations between these 24 sequences of HIV-1 *env* gene.

Table 1. Nonsynonymous (d_N) synonymous (d_S) substitution rates along branches of the tree of figure 6, estimated by our method from 24 sequences of HIV-1 env gene

	d_N	d_S	d_N/d_S		d_N	d_S	d_N/d_S
A – B	0.0102	0.1179	0.0866	D7 – D6	0.0109	0.0890	0.1225
B - C1	0.0106	0.0001	∞	D7 – D8	0.0732	0.0853	0.8581
C1 - C2	0.0213	0.0002	∞	B - E1	0.0328	0.0003	∞
C2 - C3	0.0212	0.0002	∞	E1 - E2	0.0149	0.0583	0.2555
C1 - C4	0.0106	0.0001	∞	E2 - E3	0.0105	0.2278	0.0461
C4 - C5	0.0106	0.0001	∞	E2 - E1	0.0216	0.0002	∞
D1 - D2	0.0444	0.0004	∞	E2 - E4	0.0111	0.0815	0.1362
D2 - D3	0.0327	0.0003	∞	E2 - E5	0.0327	0.0003	∞
D2 - D4	0.0000	0.0000	0.0000	E1 - E6	0.0114	0.0001	∞
D4 - D5	0.0109	0.0001	∞	E1 - E7	0.0112	0.0001	∞
D4 - D7	0.0109	0.0001	∞	E7 - E8	0.0221	0.0791	0.2794
A – F	0.0208	0.0019	10.947				

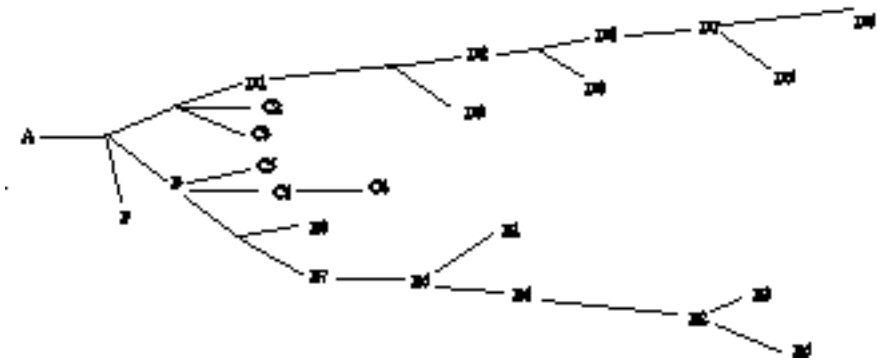

Fig. 7. *Phylogenetic tree of 24 sequential sequences of HIV-1 env gene reconstructed by maximum likelihood method (PHYLIP).*

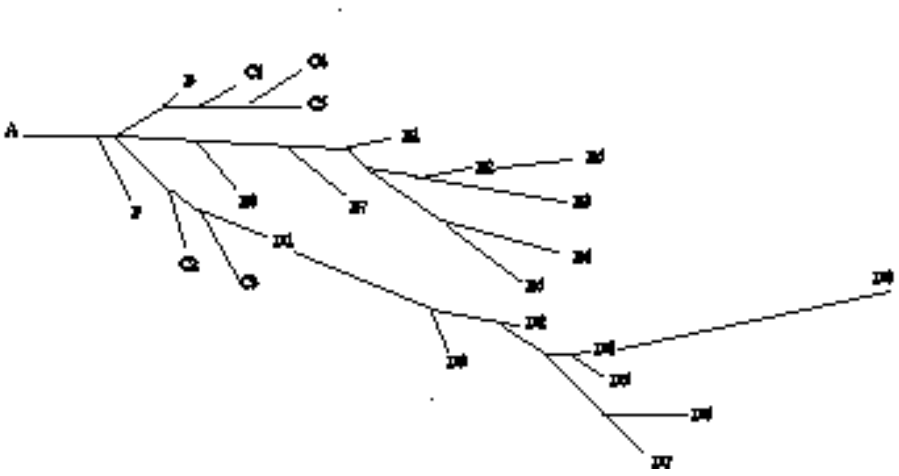

Fig. 8. *Phylogenetic tree of 24 sequential sequences of HIV-1 env gene reconstructed by neighbor-joining method (PHYLIP).*

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. Nonsynonymous (d_N) and synonymous (d_S) substitution rates along branches of the sequential phylogenetic tree are given in table 1. In figure 6, lineages under positive selection are represented by bold lines, while those under purifying selection by dotted lines. As mentioned above, if the viral sequences are under functional constraints, $d_N < d_S$, otherwise the viral sequences are under positive selection (strictly neutral evolution, $d_N = d_S$, was not found in the results).

4 Discussion

To understand viral evolution, it is important to investigate where positive selection occurs on the viral sequence. It is also very important to infer when adaptive evolution happened in the process of viral evolution. Purifying selection and adaptive evolution appear to occur alternately in the process of viral evolution, and the latter is more significant to viral evolution because the viral sequences show

wide-range divergence with accelerated evolutionary rate. The analysis of 24 sequences of HIV-1 envelope gene in this study strongly supports this hypothesis. In the beginning of infection (from year 0 to year 2), nonsynonymous substitutions were hardly observed in sequence A and the viral sequence evolved in neutral form. However, in year 3, the nonsynonymous substitutions rapidly increased and exceeded synonymous substitutions, with d_N/d_S ratio well above one. The 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 and 5; as a result, a new group, D, appeared. From year 6, however, the nonsynonymous substitution rate was rapidly reduced and the viral sequences switched to purifying selection again.

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 We thank three anonymous reviewers for giving us very helpful comments. This study was supported by a grant to women researcher to F. Ren from ESSO SEKIYU K.K. Japan.

References

[1] M. Kimura, “Evolutionary rate at the molecular level” *Nature* **217**, 624-626 (1983)

[2] T. Gojobori, “Codon substitution in evolution and the saturation of synonymous changes” *Genetics* **105**, 1011-1027 (1983)

[3] M. Nei and T. Gojobori, “Simple methods for estimating the numbers of synonymous and nonsynonymous substitutions” *Mol. Biol. Evol.* **3**, 418- 426 (1986)

[4] N. Goldman and Z. Yang, “A codon-based model of nucleotide substitution for protein-coding DNA sequence” *Mol. Biol. Evol.* **11,** 725-736 (1994)

[5] Z. Yang and R. Nielsen, "Synonymous and nonsynonymous rate variation in nuclear genes of mammals" *J. Mol. Evol.* **46**: 409-418 (1998)
[6] 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* **89,** 4835-4839 (1992)
[7] Z. Yang, "PAML: A program for package for phylogenetic analysis by maximum likelihood" *CABIOS* **15**, 555-556 (1997)
[8] J. Felsenstein, "Evolutionary trees from DNA sequences: a maximum likelihood approach" *J. Mol. Evol*. 17:368-376 (1981)
[9] N. Saitou and M. Nei, "The neighbor-joining method: a new method for reconstructing phylogenetic trees" *Mol. Biol. Evol*. **4**, 406-425 (1987)
[10] 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, 326-337 (1999)