Proc. Natl. Acad. Sci. USA
Vol. 85, pp. 2469-2473, April 1988
Biochemistry

Sequence comparisons of retroviral proteins: Relative rates of change and general phylogeny

(sequence comparisons/retrovirus relationships/recombination)

M. A. McCLURE, M. S. JOHNSON, D.-F. FENG, AND R. F. DOOLITTLE

Center for Molecular Genetics and Department of Chemistry, University of California at San Diego, La Jolla, CA 92093

Contributed by R. F. Doolittle, December 15, 1987

ABSTRACT **The inferred amino acid sequences of 10 specific gene products from nine retroviruses were aligned by computer, all evolutionary distances between them calculated, and evolutionary trees constructed. Not unexpectedly, the various gene products are changing at different rates, the reverse transcriptase being the least and the envelope proteins the most different from one retrovirus to another. For the most part, trees based on the retroviral enzyme sequences are congruent, indicating that extensive genetic recombination has not been a major factor in the evolution of the central part of the genome. In the case of envelope protein sequences, however, the sequences clearly exhibit evidence of multiple crossover events between quite distantly related retroviruses. A composite phylogenetic tree was constructed from the four retroviral enzyme sequences, and a number of important historical happenings were interpreted in the light of the time scale it affords.**

In recent years the complete nucleotide sequences of a number of retroviruses have been published, presenting an opportunity for a detailed analysis of their evolutionary relationships. This is an especially interesting area of inquiry because of the very high mutation rate observed in retroviruses (1, 2), on the one hand, and reports of widespread genetic recombination (3–8), on the other. There is also some confusion about evolutionary relationships among the retroviruses, at least partly because of the limitations of taxonomic criteria set in place before sequence data became available. (9).

Although a number of studies have already appeared that include sequence-based phylogenetic trees for various retroviruses (10–14), we felt that not all issues had been fully addressed in those reports, and, in some cases, the interpretation of events was arguable. Accordingly, we undertook an analysis of a large set of retroviral sequences with an eye to answering three questions: (*i*) how fast are the various gene products of retroviruses changing relative to each other, (*ii*) has recombination been a prominent factor in the evolution of these retroviruses, (*iii*) what is the relationship of these viruses one to another? We are not concerned here with those recombinational events involving host genes that gave rise to the acutely transforming viruses.

Our approach to measuring the rates of change of the various gene products was straightforward. First, we determined the degree of sequence difference for a given gene product between all pairs of the viruses under study, in each case normalizing the value to the difference observed between the reverse transcriptase sequences of the same two viruses. This way we dealt with relatively large numbers of data that could be averaged and the amount of variation assessed, a factor that bears significantly on the reliability of sequence-based trees. The tactic would only yield reliable results, of course, for situations in which genetic recombination was not a significant factor. It was anticipated that recombinational events, if they had occurred, would be readily detected when phylogenetic trees were constructed from the distance values. If anomalies were found, as indeed they were, it would be necessary to omit those particular data from the rate-measurement averaging.

In the analyses of the nine widely differing retroviruses, most of their gene products yielded closely congruent trees, but an anomalous pattern emerged from the comparisons of the transmembrane proteins. Because previous reference had been made to the possibility of recombination in the envelope proteins of some of the retroviruses in that set of nine, and because the results we observed were somewhat at variance with other interpretations involving several other retroviruses, we expanded our study to include an additional eight retroviruses. Although not all of these viruses had been completely sequenced, enough data were available to provide a coherent picture of several major evolutionary events.

METHODS

All computer operations were conducted on a DEC 11/730 VAX computer running the UNIX (Berkeley 4.3) operating system. Retrovirus sequences were taken from the original literature (6, 8, 13–28). Alignments were obtained by a progressive scheme designed to emphasize historical connections (29), and evolutionary trees were constructed on the basis of these alignments. In essence, the progressive alignment method begins with a conventional Needleman–Wunsch alignment (30) of the two most similar sequences of a set and then serially adds the next most similar sequence or set of sequences. Gaps are preserved in all successive alignments, and residue scoring by the minimum mutation matrix (31) is averaged across the entire existing set of sequences. Pairwise evolutionary distances (*D*) were determined after progressive alignment from the relationship

$$D = -\ln\frac{(S_{\text{real}} - S_{\text{random}})}{(S_{\text{identical}} - S_{\text{random}})} \times 100.$$

The *D* values were then used in a tree-growing program similar to that described by Fitch and Margoliash (32). In the case of a composite tree involving the four retroviral enzymes, *D* values were taken from each individual set and combined in proportion to their average lengths.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: RNP, ribonucleoprotein; BaEV, baboon endogenous virus; BLV, bovine leukemia virus; HIV, human immunodeficiency virus; HTLV, human T-cell leukemia virus; MMTV, mouse mammary tumor virus; Mo-MLV, Moloney murine leukemia virus; MPMV, Mason–Pfizer monkey virus; REV, reticuloendotheliosis virus; RSV, Rous sarcoma virus; SMRV, squirrel monkey retrovirus; SRV, simian retrovirus.

RESULTS

We began our analysis by delineating the inferred protein sequences of each of the mature gene products common to nine widely divergent retroviruses, including the avian Rous sarcoma virus (RSV), human T-cell leukemia viruses I (HTLV-I) and II (HTLV-II), bovine leukemia virus (BLV), simian retrovirus I (SRV-I), mouse mammary tumor virus (MMTV), visna lentivirus of sheep (Visna), human immunodeficiency virus (HIV), and Moloney murine leukemia virus (Mo-MLV). Where possible, the amino and carboxyl termini for the putative gene products were set on the basis of published experiments as determined by actual analysis of the viral proteins (Fig. 1). Often, however, it was necessary to make an arbitrary decision on the basis of terminator codons or alignments with other sequences whose termini had been experimentally determined. The situation is complicated because many of the genes occur in overlapping and out-of-step reading frames (34).

Altogether, 10 different segments were defined for study. These included three regions of the *gag*-encoded structural proteins denoted, respectively, *gag* amino for a segment near the amino terminus, *gag* core for the adjoining segment that spans approximately 225 residues, and *gag* RNP (ribonucleoprotein) for the approximately 80-residue segment that binds to RNA. The remaining segments analyzed were, in order of their occurrence on the genome, the protease, the reverse transcriptase, a section we call the "tether" (35), the ribonuclease H, the endonuclease, and, finally, the "outer" and "inner" portions of the envelope protein, the latter actually being the transmembrane protein.

Relative Rates of Change. The relative rates of change of the 10 defined gene products were measured by normalizing every postalignment pairwise distance to the corresponding value observed for the reverse transcriptase. As expected, the rates of change differ significantly. With the exception of the relatively short *gag* RNP segment, the enzymes change more slowly than the nonenzyme proteins (Table 1). The reverse transcriptase is the slowest-changing entity, and the outer envelope and amino-terminal *gag* proteins are the fastest. The endonucleases and ribonucleases H are changing at 1.4 ± 0.2 and 1.6 ± 0.3, the rate of the reverse transcriptases, respectively. The proteases are changing 1.8 ± 0.3 times faster than the reverse transcriptases, and the tether regions more than twice as fast. The outside portions of the envelope proteins are changing the most rapidly of all (Table 1).

Evidence for Recombination. The phylogenetic trees obtained for the four enzyme sequences (reverse transcriptase, ribonuclease H, endonuclease, and protease) were, for the most part, congruent; they were also consistent with many other characteristics that can be used to group these viruses, including the general aspect of their gene arrangements shown in Fig. 1. For most of the gene products and all of the enzymes, the Mo-MLV sequences were the most distant from the others (Fig. 2). When envelope protein sequences were used to make the trees, however, a quite different picture emerged, apparently because of a major genetic recombinational event. In the transmembrane protein tree,

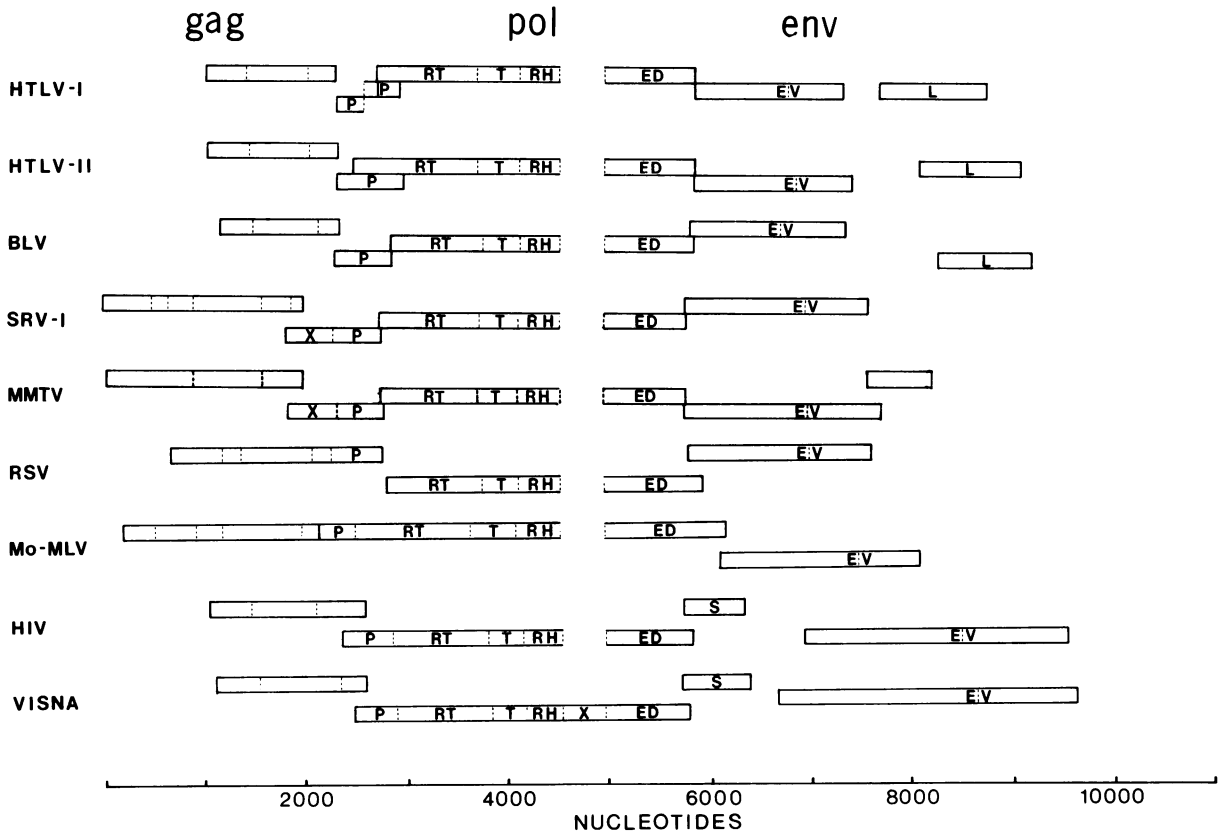

FIG. 1. Coding regions from the genomes of nine retroviruses are depicted for each of the three reading frames. The *gag* region encodes three to six different functional proteins; the *pol* region includes the protease (P), the reverse transcriptase (RT), a putative connecting segment denoted a "tether" (T), the ribonuclease H (RH), and the endonuclease (ED). The *env* region encompasses the envelope (EV) sequences. In addition, long (L) and short (S) open reading frames (HTLV-I, HTLV-II, BLV, HIV, and visna) are indicated. The segment denoted "X" corresponds to a duplication of the protease region that in some viruses has been translocated to a position between the ribonuclease H and endonuclease regions (33). Each of the boxed coding regions is oriented relative to the *gag* region arbitrarily placed in the first frame. Dashed lines represent functional boundaries; solid lines indicate beginning and ending positions of open reading frames and stop codons within a functional region of a reading frame.

Table 1. Relative rates of change of retroviral proteins

Protein*	Relative rate†
Reverse transcriptase (300)	1.0
Endonuclease (298)	1.4 ± 0.2
gag RNP (81)	1.5 ± 0.2
Ribonuclease H (133)	1.6 ± 0.3
gag core (228)	1.6 ± 0.3‡
Protease (115)	1.8 ± 0.3
Envelope, inside (194)	1.9 ± 0.3§
Tether (127)	2.2 ± 0.4
gag amino (144)	2.4 ± 0.4
Envelope, outside (284)	2.6 ± 0.5§

*Values in parentheses are average lengths in residues after arbitrary trimming to most readily aligned segments.
†Value relative to reverse transcriptase; average of 36 determinations, presented with standard deviation.
‡Based on 28 determinations (8 others were omitted because of evidence for recombination).
§Based on 20 determinations (16 others were omitted because of evidence for recombination).

for example, the Mo-MLV appears closely related to SRV-I. Conversely, in the cases of nonenvelope proteins, MMTV is close to SRV-I, but in the transmembrane protein tree it is the most distantly related to the others (Fig. 2).

The outer envelope protein sequences have changed so much that it is difficult to make a reliable tree; still the general topology obtained resembles that of the transmembrane proteins, indicating that the crossover occurred at a point such that a major portion of the envelope gene was involved in the exchange. Beyond that, hydropathy plots (36) of the transmembrane proteins of all nine viruses were consistent with an envelope exchange. In this regard, the MMTV pattern, although containing some unique features, most closely resembles those of HIV and visna. In contrast, the profiles of Mo-MLV and SRV-I are virtually indistinguishable.

Likely Gene Conversions. We uncovered two other instances in this stage of our study where recombination appears to have occurred, both of which involve the closely related retroviruses HTLV-I and HTLV-II. Thus, the envelope protein sequences in these two viruses are 85% identical, which is much too similar relative to the other gene products to be the result of ordinary divergence. The reverse transcriptase sequences of these two viruses, otherwise the most conserved of the segments under study, are only 70% identical. It appears the one of the viruses has captured the other's envelope protein gene during the interval since the initial divergence. A similar situation exists for the gag core proteins of these two viruses, which are also more than 85% identical. The high degrees of similarity appear to be the result of some kind of copy-choice having occurred during a period of coinfection by these two human viruses.

Composite Tree. After the completion of our study on the nine retroviruses discussed above, a number of other retroviral sequences became available, and it was clear that some of these would be useful in sorting out apparently anomalous resemblances. Accordingly, we constructed a composite tree from the four sets of enzyme sequences from the original set of nine viruses and then added data from a number of other retroviruses. Full sequences were not available in all cases, but we used our table of rates of change (Table 1) to scale the relative position for each new member. For example, the baboon endogenous virus (BaEV) was added to the tree on the basis of its *gag* protein sequences alone (26), after suitable scaling from the rates in Table 1. Also, we limited our alignment of endonuclease segments to approximately 180 residues, because that was

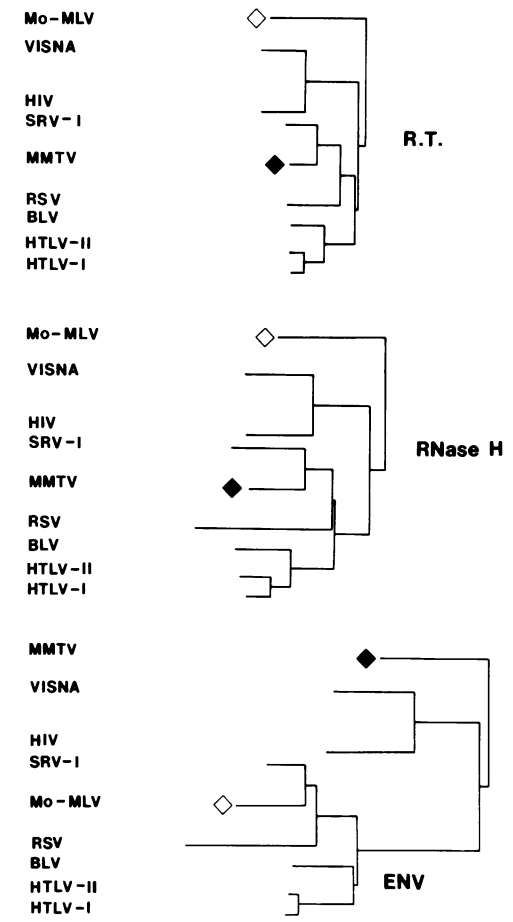

FIG. 2. Evidence for recombination between progenitors of MMTV and Mo-MLV is revealed by differences in evolutionary trees for different gene products. In the cases of reverse transcriptase (R.T.) and ribonuclease H (RNase H), the Mo-MLV is the "outlier," but in the case of the transmembrane envelope protein tree (ENV), Mo-MLV has exchanged positions with MMTV.

all the sequence available for the squirrel monkey retrovirus (SMRV) enzyme (6).

The composite tree so obtained was consistent with simple explanations for some genetic anomalies but not others. For example, one genetic event accounts for one group of retroviruses having a "single-RNP" and all the others having a duplicated version (37). Similarly, a single duplication event is consistent with the distribution of duplicated proteases (19) in the hamster intracisternal A-type particle (IAP-H18), the human endogenous retrovirus (HERV), MMTV, Mason–Pfizer monkey virus (MPMV), SR-I, and—in all likelihood—SMRV, that part of the last-named not yet having been reported. Another single event allows the translocation (33) of the duplicated segment to a different place in a line leading to three other retroviruses (Fig. 3).

The anomalous envelope pattern revealed in the various trees involving MMTV and Mo-MLV (Fig. 2) was not so easily explained, however. Thus, the human endogenous retrovirus (HERV) has an envelope sequence that resembles that of MMTV, the anomalous outlier in Fig. 2. At the same time, three other retroviruses in this cluster (SRV-I, MPMV, and SMRV) have envelope sequences that are ordinary and in-line with a direct descendancy. At the least, two recombinational events are required to explain this distribution, given the branching order based on our alignment of enzyme sequences.

The inclusion in the tree of a second avian retrovirus, the reticuloendotheliosis virus (REV-A), was also revealing.

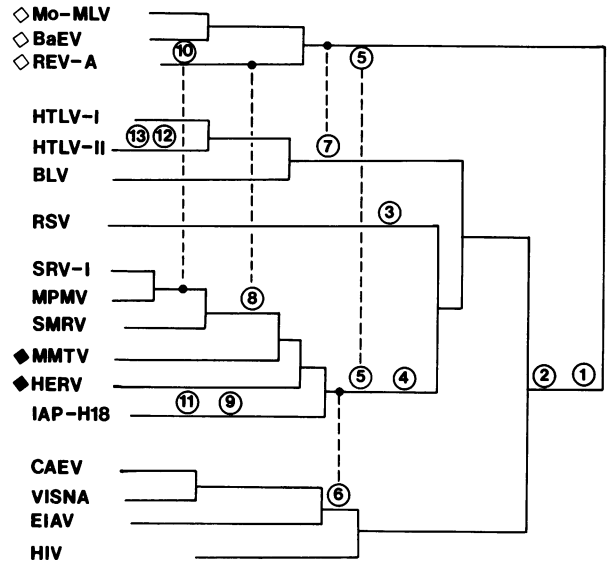

FIG. 3. Composite phylogenetic tree of 17 retroviruses. An initial tree was constructed on the basis of the four enzyme sequences from 12 retroviruses. Five additional members (CAEV, SMRV, MMTV, REV-A, and BaEV) were then added on the basis of scaled similarities for various gene products, using the relative rates listed in Table 1. The circled numbers refer to the following historical events: 1, duplicate RNP; 2, primer-binding-site change Pro→Lys; 3, primer-binding-site change Lys→Trp; 4, duplicate protease; 5, exchange envelopes; 6, capture pseudoprotease; 7, capture Pro primer-binding site; 8, capture envelope; 9, lose envelope; 10, recapture envelope; 11, primer-binding-site change Lys→Phe; 12, capture envelope; 13, capture gag amino. Diamonds (◇, ◆) denote retroviruses with anomalous envelope proteins. Dashed lines denote horizontal gene transfers in the direction of the circled numbers. BaEV, baboon endogenous virus; REV-A, avian reticuloendotheliosis virus; MPMV, Mason–Pfizer monkey virus; SMRV, squirrel monkey retrovirus; HERV, human endogenous retrovirus; IAP-H18, hamster intracisternal A-type particle; CAEV, caprine arthritis/encephalitis virus; EIAV, equine infectious anemia virus.

Indeed, the varying degrees of resemblance of BaEV, REV-A, and Mo-MLV envelope sequences with those of SRV-I, MPMV, and SMRV are best explained by a series of recombinations between progenitors of these viruses. The relative timing of these events can be estimated from their dissimilarities. In this regard, the SRV-I transmembrane sequence has about 85%, 60%, and 40% identity with the corresponding sequences from BaEV, REV-A, and Mo-MLV, respectively, a reflection of three events increasingly distant in time.

DISCUSSION

Rates of Change. Although the inferred amino acid sequences of proteins with the same function in various retroviruses are recognizably related, they are also vastly different one to another. As an example, the reverse transcriptases ranged from approximately 70% identity for the closely related viruses HTLV-I and -II to less than 25% identity for visna and Mo-MLV. In the cases of the proteases, which are known to be members of the acid protease family (38, 39), the sequences from HIV and visna are as different from each other as the corresponding sequences of acid proteases from fungi and mammals (30% identical), and those of HTLV-I and HIV are as divergent as a typical mammalian acid protease and a hypothetical prokaryotic single-domained ancestor (≈ 20% identical). Thus, the variation among these nine retroviral protein sequences covers the complete span of diversity observed in the entire non-viral world. If we were to scale this change in absolute terms, using rates of change reported from direct observation (1, 2), we would conclude that all these viruses have descended from a common ancestor in the very recent past. Given the episodic nature of viral change, however, it is difficult to gauge how recent the "recent past" may be.

That the outer-envelope proteins are the fastest-changing is hardly surprising, since it has been reported that two different HIV isolates may be as much as 20% different in these proteins (40). Actually, we were surprised that the rate of change of outer-envelope proteins was not even faster, since *a priori* we would have expected a higher proportion of adaptive (as opposed to neutral) changes in these proteins. It is likely that constraints exist for certain features, including the regions involved in the disulfide bonding of the outer and the transmembrane segments (41). Still, exposed as they are, they exhibit very rapid evolutionary change, presumably as a natural adaptation for evading host immunologic defense systems (42–44).

Recombinational Events. In the past, recombination has often been raised as a factor in the evolution and diversity of retroviruses. Clark and Mak (3) provided compelling evidence that the Friend spleen focus-forming virus (SFFV) is doubly recombinant, its 5′, central, and envelope portions being derived from three different viruses. They also showed that the envelope region had recombined at different points in various strains. In another case (7), the possibility was raised, on the basis of fragmentary sequence data, that the HIV envelope protein is a product of a recombination with MMTV. It has been reported, also, that MPMV is the product of a recombination, its envelope protein being anomalously similar to that of the avian REV (8). Other workers (6), using hybridization methods, have noted an inconsistency with regard to the relationships of envelope proteins relative to the rest of the viral genome, and the anomalous similarity of the B-type MMTV envelope protein sequence to those of typical C-type retrovirus envelopes has been remarked on elsewhere (6, 45).

The composite phylogenetic tree presented here (Fig. 3) puts many of these observations into perspective. In fact, we would contend that neither HIV nor MPMV is the direct result of a recombinational event. Rather, in both cases past comparisons were made inadvertently with other viruses that *were* the result of recombination. It seems clear that one major recombinational event involved the exchange of envelope gene sequences between progenitors of two quite distantly related groups. Subsequent events involved one-way horizontal transfers that we have dubbed "capture events" in Fig. 3 to distinguish them from the two-way exchange.

Primer-Binding Sites. One characteristic that has been used to categorize retroviruses in the past has been the nature of their primer-binding sites. Thus, all retroviruses have segments appearing at their genomic termini denoted "long terminal repeats" (LTRs) that contain an 18-base-pair section that is complementary to a particular host-cell tRNA. The amino acid corresponding to that tRNA is often, but not always, the same for closely related retroviruses (46). For example, most mammalian C-type retroviruses have proline-tRNA binding sites.

Examination of the 18-base-pair primer-binding site in the various retroviruses represented in our composite tree revealed the likelihood of a horizontal transfer of a primer-binding site from one group to another (Fig. 3). In this regard, the entire 18-base-pair primer-binding site of the HTLV-I/HTLV-II/BLV group is identical to that found in the Mo-MLV group, all corresponding to the proline tRNA. Since all of the other 11 retroviruses have primer-binding sites that differ one to another, it seems almost certain that the two groups could not have maintained or established these identities by any means other than horizontal transfer.

It is likely that other captures or exchanges of primer-binding sites have occurred, although we have not tried to position them on the tree. For example, the 18-base-pair primer-binding site of visna has no fewer than 11 differences when compared with HIV, but it has only 2 differences when compared with SRV-I.

Taxonomy, Phylogeny, and Evolution. The overall phylogeny of retroviruses put forth here was established on the basis of the inferred amino acid sequences of their four enzymes (Fig. 3). The tree is concordant with the existence of various special features, including specific open reading frames peculiar to various subgroups and several internally duplicated or translocated segments. It divides contemporary retroviruses into five groups, including two distinct and distantly related families of leukemia viruses composed of traditional C-type retroviruses on the one hand (9), and E-type on the other (13). The third group is typified by the avian RSV. A fourth group consists of morphologically diverse members, including an intracisternal A-type particle (IAP-H18), a B-type retrovirus (MMTV), and several D-type retroviruses. The fifth is a group of slow viruses that includes HIV, visna lentivirus, caprine arthritis–encephalitis virus (CAEV), and equine infectious anemia virus (EIAV).

Although recombination has occurred between some ancestors of these groups, the extent has not been great enough to confound a tracing of the basic evolutionary network. Instead, like other special features, the products of recombination can serve as directional markers for the path of evolution. Both residue replacement and recombination are more frequent in the distal regions of retroviruses. It is likely, of course, that recombination occurs at a constant frequency along the entire retroviral genome, but those events that survive natural selection occur more often in the outlying portions. In this regard, the 4-kilobase section encoding the four enzymes appears more refractory to both recombination and residue substitution than either the envelope or the terminal gag regions.

This work was supported by grants from the American Cancer Society and by National Institutes of Health Grant GM34434.

1. Gojobori, T. & Yokoyama, S. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 4198–4201.
2. Yokoyama, S. & Gojobori, T. (1987) *J. Mol. Evol.* **24,** 330–336.
3. Clark, S. P. & Mak, T. W. (1984) *J. Virol.* **50,** 759–765.
4. Troxler, D. H., Boyars, J. K., Parks, W. P. & Scolnick, E. M. (1977) *J. Virol.* **22,** 361–372.
5. Troxler, D. H., Lowy, D., Howk, R., Young, H. & Scolnick, E. M. (1977) *Proc. Natl. Acad. Sci. USA* **74,** 4671–4675.
6. Chiu, I.-M., Callahan, R., Tronick, S. R., Schlom, J. & Aaronson, S. A. (1984) *Science* **223,** 364–370.
7. Toh, H. & Miyata, T. (1985) *Nature (London)* **316,** 21–22.
8. Sonigo, P., Barker, C., Hunter, E. & Wain-Hobson, S. (1986) *Cell* **45,** 375–385.
9. Teich, N. (1985) in *RNA Tumor Viruses, Molecular Biology of Tumor Viruses*, eds. Weiss, R., Teich, N., Varmus, H. & Coffin, J. (Cold Spring Harbor Lab., Cold Spring Harbor, NY), 2nd Ed., Part 2, pp. 1–16.
10. Wain-Hobson, S., Alizon, M. & Montagnier, L. (1985) *Nature (London)* **313,** 743.
11. Chiu, I.-M., Yaniv, A., Dahlberg, J. E., Gazit, A., Skuntz, S. F., Tronick, S. R. & Aaronson, S. A. (1985) *Nature (London)* **317,** 366–368.
12. Gonda, M. A., Braun, M. J., Clements, J. E., Pyper, J. M., Wong-Staal, F., Gallo, R. C. & Gilden, R. V. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 4007–4011.
13. Sagata, N., Yasunaga, T., Tsuzuku-Kawamura, J., Ohishi, K., Ogawa, Y. & Ikawa, Y. I. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 677–681.
14. Ono, M., Yasunaga, T., Miyata, T. & Ushikubo, H. (1986) *J. Virol.* **60,** 589–598.
15. Schwartz, D. E., Tizard, R. & Gilbert, W. (1983) *Cell* **32,** 853–869.
16. Seiki, M., Hattori, S., Hirayama, Y. & Yoshida, M. (1983) *Proc. Natl. Acad. Sci. USA* **80,** 3618–3622.
17. Shimotohno, K., Takahashi, Y., Shimizu, N., Gojobori, T., Golde, D. W., Chen, I. S. Y., Miwa, M. & Sugimura, T. (1985) *Proc. Natl. Acad. Sci. USA* **82,** 3101–3105.
18. Rice, N. R., Stephens, R. M., Burny, A. & Gilden, R. V. (1985) *Virology* **142,** 357–377.
19. Power, M. D., Marx, P. A., Bryant, M. L., Gardner, M. B., Barr, P. J. & Luciw, P. A. (1986) *Science* **231,** 1567–1572.
20. Moore, R., Dixon, M., Smith, R., Peters, G. & Dickson, C. (1987) *J. Virol.* **61,** 480–490.
21. Sonigo, P., Alizon, M., Staskus, K., Klatzmann, D., Cole, S., Danos, O., Retzel, E., Tiollais, P., Haase, A. & Wain-Hobson, S. (1985) *Cell* **42,** 369–382.
22. Ratner, L., Haseltine, W., Patarca, R., Livak, K. J., Starcich, B., Josephs, S. F., Doran, E. R., Rafalski, J. A., Whitehorn, E. A., Baumeister, K., Ivanoff, L., Pettewat, S. R., Jr., Pearson, M. L., Lautenberger, J. A., Papas, T. S., Ghrayeb, J., Chang, N. T., Gallo, R. C.& Wong-Staal, F. (1985) *Nature (London)* **313,** 277–284.
23. Shinnick, T. M., Lerner, R. A. & Sutcliffe, J. G. (1981) *Nature (London)* **293,** 543–548.
24. Yaniv, A., Dahlberg, J., Tronick, S. R., Chiu, I.-M., Aaronson, S. A. (1985) *Virology* **145,** 340–345.
25. Stephens, R. M., Casey, J. W. & Rice, N. R. (1986) *Science* **231,** 589–594.
26. Tamura, T.-A. (1983) *J. Virol.* **47,** 137–145.
27. Wilhelmsen, K. C., Eggleton, K. & Temin, H. (1984) *J. Virol.* **52,** 172–182.
28. Ono, M., Toh, H., Miyata, T. & Awaya, T. (1985) *J. Virol.* **55,** 387–394.
29. Feng, D.-F. & Doolittle, R. F. (1987) *J. Mol. Evol.* **25,** 351–360.
30. Needleman, S. B. & Wunsch, C. D. (1970) *J. Mol. Biol.* **48,** 443–453.
31. Dayhoff, M. O., ed. (1978) *Atlas of Protein Sequence and Structure* (Natl. Biomed. Res. Found., Washington, DC), Vol. 5, Suppl. 2, pp. 353–358.
32. Fitch, W. M. & Margoliash, E. (1967) *Science* **15,** 279–284.
33. McClure, M. A., Johnson, M. S. & Doolittle, R. F. (1987) *Proc. Natl. Acad. Sci. USA* **84,** 2693–2697.
34. Levin, J. G., Her, S. C., Rein, A., Messer, L. I. & Gerwin, B. I. (1984) *J. Virol.* **51,** 470–478.
35. Johnson, M. S., McClure, M. A., Feng, D.-F., Gray, J. & Doolittle, R. F. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 7648–7652.
36. Kyte, J. & Doolittle, R. F. (1982) *J. Mol. Biol.* **157,** 105–132.
37. Covey, S. N. (1986) *Nucleic Acids Res.* **14,** 623–633.
38. Toh, H., Ono, M., Saigo, K. & Miyata, T. (1985) *Nature (London)* **315,** 691.
39. Pearl, L. H. & Taylor, W. R. (1987) *Nature (London)* **329,** 351.
40. Desai, S. M., Kalyanaraman, V. S., Casey, J. M., Srinivasaw, A., Andersen, P. P. & Devare, S. G. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 8380–8384.
41. Starcich, B. R., Hahn, B. H., Shaw, G. M., McNeely, P. D., Modrow, S., Wolf, H., Parks, E. S., Parks, W. P., Josephs, S. F., Gallo, R. C. & Wong,-Staal, F. (1986) *Cell* **45,** 637–648.
42. Alizon, M., Wain-Hobson, S., Montagnier, L. & Sonigo, P. (1986) *Cell* **46,** 63–74.
43. Coffin, J. M. (1986) *Cell* **46,** 1–4.
44. Wong-Staal, F. & Gallo, R. C. (1985) *Nature (London)* **317,** 395–403.
45. Thayer, R. M., Power, M. D., Bryant, M. L., Gardner, M. B., Barr, P. J. & Luciw, P. A. (1987) *Virology* **157,** 317–329.
46. Chen, H. R. & Barker, W. C. (1984) *Nucleic Acids Res.* **12,** 1767–1778.