29 Bork, P. and Koonin, E. V. (1998) Nature Genet. **18**, 313–318
30 Martin, A. C., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Lawskowski, R. A., Mitchell, J. B., Taroni, C. and Thornton, J. M. (1998) Structure **6**, 875–884
31 Kraulis, P. J. (1991) J. Appl. Cryst. **24**, 946–950

Received 9 September 1999

Origins and evolution of AIDS viruses: estimating the time-scale

P. M. Sharp*[1], E. Bailes*, F. Gao†, B. E. Beer‡, V. M. Hirsch‡ and B. H. Hahn†
*Institute of Genetics, University of Nottingham, Queens Medical Centre, Nottingham NG7 2UH, U.K., †Departments of Medicine and Microbiology, University of Alabama at Birmingham, Birmingham, Alabama 35294, U.S.A., and ‡Laboratory of Molecular Microbiology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Rockville, Maryland 20852, U.S.A.

Abstract

The primate lentiviruses comprise SIV strains from various host species, as well as two viruses, HIV-1 and HIV-2, that cause AIDS in humans. The origins of HIV-1 and HIV-2 have been traced to cross-species transmissions from chimpanzees and sooty mangabey monkeys respectively. Two approaches have been taken to estimate the time-scale of the evolution of these viruses. Certain groups of SIV strains appear to have evolved in a host-dependent manner, implying a time-scale of many thousands or even millions of years. In stark contrast, molecular clock calculations have previously been used to estimate a time-scale of only tens or hundreds of years. Those calculations largely ignored heterogeneity of evolutionary rates across different sites within sequences. In fact, the distribution of rates at different sites seems extremely skewed in HIV-1, and so the time-depth of the primate lentivirus evolutionary tree may have been underestimated by at least a factor of ten. However, these date estimates still seem to be far too recent to be consistent with host-dependent evolution.

Introduction

Two distinct retroviruses, human immunodeficiency virus types 1 and 2 (HIV-1 and HIV-2), cause the acquired immune deficiency syndrome (AIDS). Together with related simian immunodeficiency viruses (SIVs), found in other primate species, these comprise the primate lentiviruses. These viruses exhibit an extraordinary degree of sequence diversity and provide an interesting and challenging model system in which to study molecular evolution. Furthermore, the results obtained provide insights into the origins of AIDS, and have important practical implications for understanding and tackling the ongoing AIDS pandemic.

Origins of AIDS viruses

Lentiviruses have been isolated from numerous primate species. Phylogenetic analyses indicate that these viruses fall into five major, approximately equidistant, lineages (Figure 1). The two viruses causing AIDS in humans belong to different lineages and represent recent cross-species transmissions from different sources. The vast majority of cases of HIV infection worldwide are due to strains of HIV-1. These viruses cluster with $\mathrm{SIV_{CPZ}}$ from chimpanzees (*Pan troglodytes*). Although $\mathrm{SIV_{CPZ}}$ was first reported in 1990 [1], there remained the question of whether chimpanzees had, like humans, recently acquired the virus from some other species [2,3]. However, we have recently presented evidence strongly suggesting that one particular subspecies of chimpanzee, *P. t. troglodytes*, does indeed constitute the source of HIV-1 [4].

HIV-2 is only common in west Africa. Strains of HIV-2 cluster with $\mathrm{SIV_{SM}}$ from sooty mangabeys (*Cercocebus atys*), and with $\mathrm{SIV_{MAC}}$ from macaques (Figure 1). The macaque viruses were among the first SIV strains to be characterized, but it was soon realized that macaques in the wild are not infected with SIV, and that all of the isolates reflect transmission in captivity. In contrast, sooty mangabeys have been shown to be naturally infected with SIV. These monkeys inhabit west Africa, and represent the source of HIV-2.

Both HIV-1 and HIV-2 exhibit considerable genetic diversity. HIV-1 strains have been classified into three groups (M, N and O), with the M group further divided into numerous subtypes (A–J, so far). The phylogenetic interspersion of the $\mathrm{SIV_{CPZ}}$ and HIV-1 lineages indicates that the

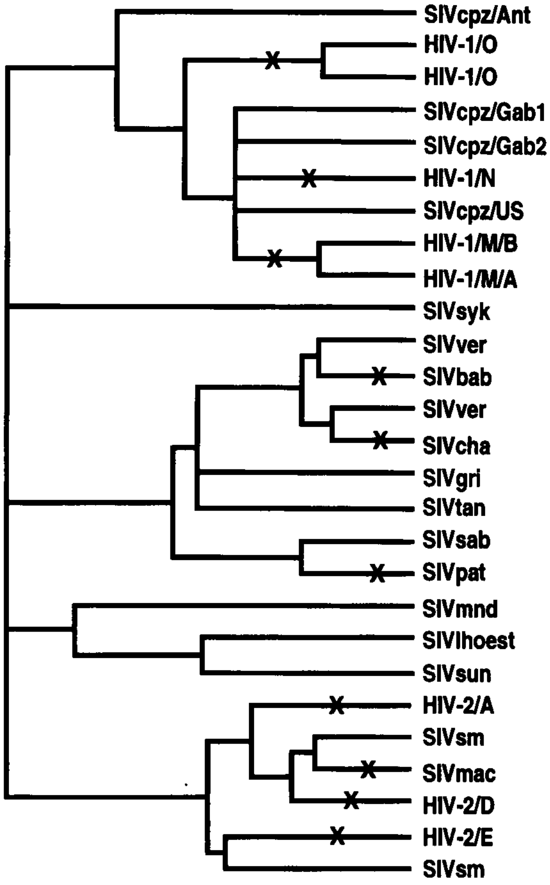
Key words: gamma distribution, HIV-1, host-dependent evolution, molecular clock, primate lentiviruses.

[1] To whom correspondence should be addressed.

Figure 1

Phylogenetic relationships among representative primate lentiviruses

For HIV-1, the three groups (M, N and O) are shown, as well as two of the ten described subtypes of HIV-1 group M. For HIV-2, three of the six known subtypes of HIV-2 are represented. Strains of SIV have a subscript denoting the species from which they were isolated (see text for details). Probable positions of cross-species transmission are indicated by 'X'. Figure generated from data in [2,4,34,46].



three groups must each have arisen through separate transmissions from chimpanzees [4]. HIV-2 strains have been classified into six subtypes (A–F); again, the relationships among these subtypes and the known examples of $\mathrm{SIV_{SM}}$ indicate that there must have been at least four cross-species transmissions, and it is possible that all six subtypes arose independently [5,6].

The three other major lineages of primate lentivirus (Figure 1) represent natural infections of African green monkeys ($\mathrm{SIV_{AGM}}$, from the *Chlorocebus* genus), of Sykes monkeys ($\mathrm{SIV_{SYK}}$, from *Cercopithecus albogularis*), and of l'Hoest and sun-tailed monkeys ($\mathrm{SIV_{LHOEST}}$ and $\mathrm{SIV_{SUN}}$ from the *Cercopithecus lhoesti* group). Within the $\mathrm{SIV_{AGM}}$ clade there are additional examples of cross-species transmission. A yellow baboon (*Papio cynocephalus*) from Tanzania [7], a chacma baboon (*Papio ursinus*) from South Africa [8], and a Patas monkey (*Erythrocebus patas*) from Senegal [9] were each found to have become naturally infected with a virus closely related to those from the local species of African green monkey. Further examples of SIV from other species have been reported recently, and are in the process of being fully characterized. These include SIVs from red-capped mangabeys (*Cercocebus torquatus*) [10], drills (*Mandrillus leucophaeus*) [11] and talapoins (*Myopithecus talapoin*) [12].

Thus phylogenetic analyses have shown that there must have been numerous instances of cross-species transmission of primate lentiviruses, not only simian-to-human, but also simian-to-simian. One outstanding question is: when did these evolutionary events occur?

Time-scale?

In the absence of a fossil record, two main approaches have been used in attempts to establish time-scales for viral evolution. For rapidly evolving viruses, it is possible to calibrate their rate of evolution by comparison of viruses isolated at different time points, and then to use this molecular clock to estimate the dates of divergences within the evolutionary tree. This procedure has been used, for example, with influenza A viruses [13]. The other approach has been to find instances where viruses appear to have been evolving in a host-dependent manner, i.e. without transmission between different species of host. In such cases, the divergence times of hosts also pertain to their viruses. It has been suggested that this method can be applied to herpesviruses [14].

Both techniques have been adopted to estimate the time-scale of primate lentivirus evolution. Some of the earliest attempts, in 1988, yielded dramatically discordant results. One attempt to use a molecular clock estimated that the common ancestor of HIV-1 and HIV-2 existed as recently as 1951 [15]. At the same time, after comparison of HIV-1 with an SIV from an African green monkey, others suggested that these viruses may have diverged 'gradually in concert with the evolution of primates' [16], and that it was possible that a common ancestor of these viruses infected the common ancestor of apes and Old World monkeys [17]; that ancestor is thought to have lived at least 25 million years ago. Since HIV-1, HIV-2 and $\mathrm{SIV_{AGM}}$ each belong to different major lineages of primate lentivirus (Figure 1), these two estimates pertain to the same common ancestor of all the primate lentiviruses. Thus the two different approaches led to estimates with a discrepancy of about six orders of magnitude!

Molecular clock?

We first estimated the rate of evolution of HIV-1 by comparing the distances, from an outgroup, of pairs of viruses that had been isolated at different times [18]. The latter isolates were expected (and found) to have diverged more from the outgroup sequence, and the additional extent of divergence was divided by the difference in time of isolation. The overall rate obtained was about 5×10^{-3} nucleotide substitutions per site per year, equivalent to a rate of divergence between two sequences of around 1 % per annum. Non-synonymous nucleotide substitution rates were about 1.6×10^{-3} for *gag* and *pol*, and 5.1×10^{-3} for *env*, reflecting the different rates at which these HIV-1 proteins evolve. Synonymous nucleotide substitution rates were similar in different genes, averaging about 10×10^{-3} substitutions per site per year.

Although very few sequences were available for study at that time (1987), two considerations suggest that these estimates were quite accurate. First, several subsequent, independent studies of different data sets derived very similar estimates. For example, Gojobori et al. [19] estimated rates for 13.1×10^{-3} and 3.9×10^{-3} per site per year for synonymous and non-synonymous changes in *gag*, whereas Kelly estimated a synonymous substitution rate of 11.0×10^{-3} per site per year in the V3 region of the *env* gene [20]. Furthermore, an overall rate of divergence of about 1 % per year has commonly been found for HIV-1 sequences.

Secondly, if the majority of synonymous substitutions are selectively neutral, their rate of accumulation is expected to be the same as the mutation rate [21]. The rate of mutation has been estimated as 3.4×10^{-5} per site per replication [22], while the average number of replications per year has been estimated as around 300 [23,24]. This leads to a mutation rate of around 10.2×10^{-3} per site per year, very close to the estimate of the synonymous substitution rate.

If the molecular clock has been calibrated reasonably accurately, it can be used to estimate divergence times within the lentivirus phylogeny. On the assumption that the clock rate for HIV-1 pertained to the other lentiviruses, we used the rate of non-synonymous substitution in the *pol* gene to estimate the time of the common ancestor of HIV-1, HIV-2 and $\mathrm{SIV}_{\mathrm{AGM}}$ at about 150 years ago [25].

Host-dependent evolution?

While it is clear that there have been numerous instances of cross-species transmission during the evolution of the primate lentiviruses, nevertheless certain clades of SIV appear to have evolved with their hosts for a long period of time. In such cases, it may be possible to infer the dates of branching points within the SIV phylogeny from those within the evolutionary tree of the hosts. Within the primate lentivirus phylogeny (Figure 1), there are three possible examples of such host-dependent evolution.

The first concerns $\mathrm{SIV}_{\mathrm{AGM}}$. There are four main species of African green monkeys, inhabiting abutting ranges across sub-Saharan Africa: sabaeus monkeys in west Africa, tantalus in central Africa, grivets in east Africa, and vervets from east to southern Africa. All four species are naturally infected with $\mathrm{SIV}_{\mathrm{AGM}}$, and multiple examples of each have been at least partially characterized. In phylogenetic analyses, the viruses all cluster in a host-dependent manner [26,27]. Furthermore, among the clades of viruses, $\mathrm{SIV}_{\mathrm{AGM}}$Sab appears to be the earliest diverging lineage (Figure 1), whereas, among the monkeys, the sabaeus species appears to be the most divergent [28]. These observations are consistent with the common ancestor of the African green monkey species having been infected with the common ancestor of the $\mathrm{SIV}_{\mathrm{AGM}}$ lineages. The time of the common ancestor of the African green monkey species is not known from the fossil record. However, for both mitochondrial DNA [28] and nuclear CD4 sequence comparisons [29], sabaeus and vervet monkeys exhibit a very similar amount of sequence divergence to humans and chimpanzees. Humans and chimpanzees are estimated to have shared a common ancestor 5–10 million years ago. Rates of molecular evolution appear to have been somewhat slower in apes than in Old World monkeys [30], but this still suggests that the common ancestor of the African green monkey species existed at least 1 million years ago.

A second example concerns $\mathrm{SIV}_{\mathrm{CPZ}}$. The chimpanzee *P. troglodytes* has been classified into four geographically distinct subspecies [31]. Among the four known isolates of $\mathrm{SIV}_{\mathrm{CPZ}}$, three (Gab1, Gab2 and US) came from *P. t. troglodytes*, and one (Ant) from *P. t. schweinfurthii* [4]. Although the sample size is small, again these viruses cluster in a host-dependent manner (Figure 1), suggesting that the common ancestor of the two $\mathrm{SIV_{CPZ}}$ lineages may have infected the common ancestor of the two subspecies of chimpanzee [4]. From mitochondrial DNA comparisons, that ancestor has been estimated at 500000 years ago [32].

We have very recently described a third possible example [33,34]. Two closely related species, l'Hoest monkeys (*Cercop. lhoesti*) and sun-tailed monkeys (*Cercop. solatus*), have been found to be infected with viruses ($\mathrm{SIV_{LHOEST}}$ and $\mathrm{SIV_{SUN}}$) that are more closely related to one another than to any other SIV (Figure 1), again indicating that this may reflect host-dependent evolution. The time of the common ancestor of these monkeys is not known, but they are thought to have diverged because of vegetation changes in Africa some time within the last million years.

These putative instances of host-dependent evolution of SIV suggest divergence times within the primate lentivirus phylogeny on the order of (at least) hundreds of thousands of years. The common ancestor of all of the primate lentiviruses, i.e. at the left of Figure 1, would be even older.

Resolving the discrepancy?

These two approaches to estimating the time-scale of primate lentivirus evolution yield highly discordant results. The number of primate species now known to be infected by their own specific SIV, and the apparently host-dependent evolution of certain clades of SIV, strongly suggest that the primate lentiviruses as a whole cannot have evolved only within the last millennium. This indicates that the time-scale suggested from the molecular clock is most likely to be a gross underestimate.

If the rate of evolution for HIV-1 has been overestimated, or the rate of evolution of other lentiviruses was slower than for HIV-1, the time depth could be larger. Recently, the rate of substitution in the HIV-1 *env* gene was estimated as 1.7×10^{-3} substitutions per site per year [35], i.e. about three times slower than previous estimates. Clearly, this difference is not sufficient to explain a discrepancy of orders of magnitude. Furthermore, there is no evidence that strains of SIV evolve at a substantially lower rate than HIV-1. Indeed, retroviruses all appear to have similarly high mutation rates because of the error-prone nature of reverse transcriptase, and so lower evolutionary rates could only be achieved through a lower rate of viral replication. Even though it is apparently not pathogenic, $\mathrm{SIV_{AGM}}$ appears to have a replication rate similar to that of HIV-1 [36]. Of course, if the primate lentiviruses ever became endogenous, their rate of evolution should drop to that of the host genome, i.e. around 2×10^{-9} substitutions per site per year [30]: that could immediately explain the discrepancy, but again there is no evidence that this has occurred.

Alternatively, if the estimate of the rate of HIV-1 evolution is approximately correct, then the underestimation of the time depth must reflect an underestimation of the extent of evolutionary divergence among viruses. For all but the closest comparisons, estimates of evolutionary divergence between lentiviruses must involve an attempt to correct for multiple hits, i.e. successive substitutions superimposed at a site during the evolution of a lineage. Standard methods for making these corrections assume quite simple models of sequence evolution. If the real pattern of sequence evolution deviates from these assumptions, the true distance will generally be underestimated.

Two of the assumptions may be particularly important. The first concerns the relative frequencies of the 12 possible changes among the four nucleotides. The simplest (Jukes–Cantor) model assumes that all substitutions are equally likely, whereas it is known that this is usually not the case. In particular, it is commonly observed that transitions occur more often than transversions; the 2-parameter model (from Kimura) allows for different rates of transitions and transversions. More complex models attempt to incorporate additional parameters, allowing an increased number of different categories of rate of nucleotide replacement, but are not yet widely used.

The second assumption concerns the relative frequency of substitutions at different sites within sequences. While the difference in frequency between synonymous and non-synonymous substitutions is often recognized, it is usually implicitly assumed that non-synonymous substitutions at different sites all have the same underlying rate. In fact, consideration of almost any protein alignment indicates that different sites within a protein structure can evolve at very different rates. This problem, of heterogeneity of rates across sites, was recognized nearly 30 years ago [37], but has only recently begun to receive much attention [38]. So far, the most widely used approach is to assume that the rate heterogeneity can be approximated by a gamma distribution. The gamma distribution takes different shapes, according to the value of a single shape parameter (α): with α values >1, the distribution is approximately bell-shaped, while with α values < 1, the distribution becomes progressively more L-

Table 1

Effects of substitution model on estimates of distance and time

Corrected estimates of sequence distance, and corresponding time depth, are given for various extents of transition:transversion bias (measured by κ) and among-site rate heterogeneity (measured by the shape parameter, α, of a gamma distribution; constant refers to a model assuming rate homogeneity). Upper values are the estimated number of substitutions per site; lower values are the estimated times since divergence, assuming a rate of 10^{-3} substitutions per site per year. Pairs of values are given, calculated for observed sequence differences of 0.15 and 0.25. Thousands are indicated by ˆ3.

		α					
κ	Constant	1.0	0.5	0.2	0.1	0.06	0.04
1.0	0.17/0.30 84/152	0.19/0.38 94/188	0.21/0.47 106/235	0.31/0.99 154/495	0.62/4.25 312/2125	1.81/38.7 905/19ˆ3	7.91/758 3956/379ˆ3
2.0	0.17/0.31 84/153	0.19/0.38 94/190	0.21/0.48 107/238	0.31/1.01 156/504	0.63/4.33 317/2166	1.84/39.4 922/20ˆ3	8.06/771 4028/386ˆ3
4.0	0.17/0.31 85/156	0.19/0.39 96/197	0.22/0.50 110/250	0.33/1.08 164/538	0.68/4.65 338/2325	1.97/42.2 986/21ˆ3	8.62/826 4310/413ˆ3
7.0	0.17/0.32 86/160	0.20/0.41 99/206	0.23/0.53 114/266	0.35/1.18 174/591	0.74/5.15 368/2573	2.17/46.7 1084/23ˆ3	9.49/910 4746/455ˆ3
10.0	0.17/0.32 87/162	0.20/0.43 100/213	0.23/0.56 116/279	0.37/1.27 182/637	0.79/5.59 394/2797	2.34/50.6 1172/25ˆ3	10.3/985 5137/493ˆ3

shaped. This range of alternative shapes seems a reasonable first approximation of the distribution of rates that might be expected.

The effect on estimates of evolutionary distance (and hence divergence times) of gamma-distributed among site rate heterogeneity, and of transition: transversion bias, can be modelled using equations from Jin and Nei [39]. In Table 1, we present estimated corrected evolutionary distances for two levels of observed sequence differences (15 % and 25 %), across a range of degrees of underlying among site rate heterogeneity, and transition: transversion bias. We also present the implications in terms of time depth for one particular rate of evolution; the effect of different rates on the time depth can be simply obtained by multiplication.

Under the simplest (Jukes–Cantor) model (i.e. $\kappa = 1.0$), with no rate heterogeneity, the estimated 'corrected' distances are 0.17 and 0.30. Increasing the transition: transversion ratio (κ) has little effect on these values. Increasing the amount of rate heterogeneity (decreasing α) immediately increases the estimated distance, although for the smaller difference (0.15) the effect is minor until α becomes quite small (0.2). As the rate heterogeneity increases, the effect of higher transition :transversion bias also increases. At high values of κ, and very low values of α, the estimated distances and, correspondingly, the associated estimates of time depth, are orders of magnitude higher than those obtained with the assumption of rate homogeneity.

Estimates of sequence evolution parameters

Are the patterns of nucleotide substitution in AIDS viruses sufficiently skewed that the evolutionary distances among them have been underestimated by orders of magnitude? Leitner et al. [40] have assessed the most appropriate model of evolution for short fragments of the *gag* and *env* genes by exploiting a 'known' phylogeny of closely related HIV-1 isolates from a well-described transmission chain involving nine patients. They used a gamma distribution and derived estimates of α ranging from 0.18 (for codon position 2 in *gag*) to 0.74 (for codon position 1 in *env*).

We have followed a similar approach, but using full-length sequences of HIV-1 isolates representing a greater phylogenetic diversity. We took eight published sequences for which the branching pattern of the evolutionary tree can be taken as reliably known. The sequences represent three different clades, subtypes B, D and E of the M group. Numerous analyses have established that subtypes B and D are more closely related to each other than to other subtypes (for example, see [41]). Of three subtype B sequences included, two came from different tissues of the same individual,

Table 2

Estimates of substitution model parameters for HIV-1

Estimates of the shape parameter (α) of the gamma distribution describing among site rate heterogeneity, and of the transition:transversion bias (κ), for the three different positions within codons for the three major genes of HIV-1.

	Codon position		
	1	2	3
	α/κ	α/κ	α/κ
gag	0.22/2.1	0.16/6.2	1.52/13.0
pol	0.16/8.4	0.12/6.9	1.10/14.8
env	0.43/3.1	0.30/4.4	1.07/ 6.6

and can be assumed to be more closely related than a third isolate from an unlinked individual. Analogously, of three subtype E sequences included, two came from Thailand, and can be assumed to be more closely related than a third isolate from central Africa, because the Thai subtype E epidemic appears to have resulted from a single import [42]. The PAML program [43] was used to derive maximum likelihood estimates of the transition:transversion bias (κ), and the shape parameter (α) of the gamma distribution describing heterogeneity of rates across sites, assuming this known topology of the phylogenetic tree linking these isolates. Separate analyses were performed for *gag*, *pol* and *env* genes, and for each of the three positions within codons. The α and κ values obtained are shown in Table 2.

The α values obtained for different codon positions have the relative magnitudes position $3 \gg$ position $1 >$ position 2, for all three genes. The majority of substitutions at position 3 and a small minority of substitutions at position 1 are synonymous, whereas all substitutions at position 2 are non-synonymous. Rates of synonymous substitutions are expected to be more homogeneous across different sites within a gene, and so higher α values (indicates less heterogeneity) are expected for sites undergoing synonymous changes. The α values for first and second positions, particularly in the *gag* and *pol* genes encoding the more highly conserved proteins, are extremely low; they are similar to the lowest values reported by Yang for primate mitochondrial genes [38]. Thus the α values for HIV-1 indicate that non-synonymous changes exhibit extreme heterogeneity of rates across sites. In addition, the estimated κ values (Table 2) are generally quite high, indicating a very strong bias towards transitional substitutions in HIV-1.

Conclusions

Some years ago, Eigen and Nieselt-Struwe [44] pointed out that differences in evolutionary rates at different sites could lead to problems in estimating the age of viruses. They suggested three categories of sites; constant, variable, and hypervariable, and from analysis of *env* sequences concluded that the common ancestor of the primate lentiviruses dates back at least 600–1200 years. The estimates of α and κ described above, indicating extreme biases in substitution patterns, suggest that evolutionary distances among the more divergent lentivirus sequences may have been even more severely underestimated by the use of 'standard' multiple hit correction methods. If we focus on second codon positions within the *pol* gene (since the Pol protein is the most highly conserved), at these sites $\mathrm{SIV}_{\mathrm{AGM}}$ sequences from vervet, grivet and tantalus monkeys differ by 15–19%, while HIV-1 and HIV-2 sequences differ by 25–27%, similar to the values used in Table 1. On the basis of our early estimates [18], the rate of substitution at these sites is around 10^{-3} substitutions per site per year. Using the results from Table 2 (of an α value around 0.1 and a κ value of around 7), the estimate of the time of divergence of the $\mathrm{SIV}_{\mathrm{AGM}}$ sequences would increase from less than 100 years to more than 350 years, while that for HIV-1 and HIV-2 would increase from about 150 years to around 2500. These dates are substantially older than previous estimates, but they are still not consistent with SIVs having evolved in a host-dependent fashion for hundreds of thousands of years.

Perhaps the most interesting, but also controversial, date within the lentivirus phylogeny concerns the time of origin of the M group of HIV-1. In our first attempts to date this, we estimated the time of the common ancestor of (what are now known as) subtypes B and D at around 1969, and the common ancestor of the entire M group at around 1960 [18]. These dates were shown to be too recent by analyses of partial HIV-1 sequences from a blood sample taken in Zaire in 1959: that virus lay close to, but later than, the common ancestor of subtypes B and D [45], indicating that the M group must have originated somewhat earlier, around 1940. A similar time-scale was derived recently from analysis of a large number of *env* sequences [35]. Isolates from different subtypes of group M differ by about

5–6% at position 2 in *pol*. Applying the same approach described above, with an α value of 0.12 and a κ value of 6.9, a difference of 0.06 yields a divergence time of about 50 years. Since the viruses were isolated in the late 1980s, that places the common ancestor of group M in the late 1930s, broadly consistent with these other recent estimates.

In conclusion, previous applications of the molecular clock to the evolution of the primate lentiviruses have used oversimplified assumptions and underestimated the time depth of their evolution. The degree of underestimation has been small, but significant, for recent events, and very substantial for the earlier events. We are currently examining the extent to which the incorporation of additional parameters within the model of sequence change might increase the estimated distances and push back further the perceived time depth of the primate lentivirus phylogeny.

References

1 Huet, T., Cheynier, R., Meyerhans, A., Roelants, G. and Wain-Hobson, S. (1990) Nature (London) **345**, 356–359
2 Sharp, P. M., Robertson, D. L. and Hahn, B. H. (1995) Phil. Trans. R. Soc. Lond. Ser. B **349**, 41–47
3 Wain-Hobson, S. (1998) Nature Med. **4**, 1001–1002
4 Gao, F., Bailes, E., Robertson, D. L., Chen, Y., Rodenburg, C. M., Michael, S. F., Cummins, L. B., Arthur, L. O., Peeters, M., Shaw, G. M., Sharp, P. M., and Hahn, B. H. (1999) Nature (London) **397**, 436–441
5 Gao, F., Yue, L., Robertson, D. L., Hill, S. C., Hui, H., Biggar, R. J., Neequaye, A. E., Whelan, T. M., Ho, D. D., Shaw, G. M., Sharp, P. M. and Hahn, B. H. (1994) J. Virol. **68**, 7433–7447
6 Chen, Z., Luckay, A., Sodora, D. L., Telfer, P., Reed, P., Gettie, A., Kanu, J. M., Sadek, R. F., Yee, J., Ho, D. D., Zhang, L. and Marx, P. A. (1997) J. Virol. **71**, 3953–3960
7 Jin, M. J., Rogers, J., Phillips-Conroy, J. E., Allan, J. S., Desrosiers, R. C., Shaw, G. M., Sharp, P. M. and Hahn, B. H. (1994) J. Virol. **68**, 8454–8460
8 van Rensburg, E. J., Engelbrecht, S., Mwenda, J., Laten, J. D., Robson, B. A., Stander, T. and Chege, G. K. (1998) J. Gen. Virol. **79**, 1809–1814
9 Bibollet-Ruche, F., Galat-Luong, A., Cuny, G., Sami-Manchado, P., Galat, G., Durand, J.- P., Pourrut, X. and Veas, F. (1996) J. Gen. Virol. **77**, 773–781
10 Georges-Courbot, M. C., Lu, C. Y., Makuwa, M., Telfer, P., Onanga, R., Dubreuil, G., Chen, Z., Smith, S. M., Georges, A., Gao, F., Hahn, B. H. and Marx, P. A. (1998) J. Virol. **72**, 600–608
11 Clewley, J. P., Lewis, J. C. M., Brown, D. W. G. and Gadsby, E. L. (1998) J. Virol. **72**, 10305–10309
12 Osterhaus, A. D. M. E., Pedersen, N., van Amerongen, G., Frankenhuis, M. T., Marthas, M., Reay, E., Rose, T. M., Pamungkas, J. and Bosch, M. L. (1999) Virology **260**, 116–124
13 Fitch, W. M. (1995) Phil. Trans. R. Soc. Lond. Ser. B **349**, 93–102
14 McGeoch, D. J., Cook, S., Dolan, A., Jamieson, F. E. and Telford, E. A. R. (1995) J. Mol. Biol. **247**, 443–458
15 Smith, T. F., Srinivasan, A., Schochetman, G., Marcus, M. and Myers, G. (1988) Nature (London) **333**, 573–575
16 Fukasawa, M., Miura, T., Hasegawa, A., Morikawa, S., Tsujimoto, H., Miki, K., Kitamura, T. and Hayami, M. (1988) Nature (London) **333**, 457–461
17 Mulder, C. (1988) Nature (London) **333**, 396
18 Li, W.- H., Tanimura, M. and Sharp, P. M. (1988) Mol. Biol. Evol. **5**, 313–330
19 Gojobori, T., Moriyama, E. and Kimura, M. (1990) Proc. Natl. Acad. Sci. U.S.A. **87**, 10015–10018
20 Kelly, J. K. (1994) Genet. Res. **64**, 1–9
21 Kimura, M. (1983) The Neutral Theory of Molecular Evolution, Cambridge University Press
22 Mansky, L. M. and Temin, H. M. (1995) J. Virol. **69**, 5087–5094
23 Wei, X., Ghosh, S. K., Taylor, M. E., Johnson, V. A., Emini, E. A., Deutsch, P., Lifson, J. D., Bonhoeffer, S., Nowak, M. A., Hahn, B. H., Saag, M. S. and Shaw, G. M. (1995) Nature (London) **73**, 117–122
24 Perelson, A. S., Neumann, A. U., Markowitz, M., Leonard, J. M. and Ho, D. D. (1996) Science **271**, 1582–1586
25 Sharp, P. M. and Li, W.- H. (1988) Nature (London) **336**, 315
26 Muller, M. C., Saksena, N. K., Nerrienet, E., Chappey, C., Herve, V. M. A., Durand, J.- P., Lagal-Campodonico, P., Lang, M.- C., Digoutte, J. P., Georges, A. J., Georges-Courbot, M. C., Sonigo, P. and Barre-Sinoussi, F. (1993) J. Virol. **67**, 1227–1235
27 Jin, M. J., Hui, H., Robertson, D. L., Muller, M. C., Barre-Sinoussi, F., Hirsch, V. M., Allan, J. S., Shaw, G. M., Sharp, P. M. and Hahn, B. H. (1994) EMBO J. **13**, 2935–2947
28 van der Kuyl, A. C., Kuiken, C. L., Dekker, J. T. and Goudsmit, J. (1995) J. Mol. Evol. **40**, 173–180
29 Fomsgaard, A., Muller-Trutwin, M. C., Diop, O., Hansen, J., Mathiot, C., Corbet, S., Barre-Sinoussi, F. and Allan, J. S. (1997) J. Med. Primatol. **26**, 120–128
30 Li, W.- H., Tanimura, M. and Sharp, P. M. (1987) J. Mol. Evol. **25**, 330–342
31 Gagneux, P., Wills, C., Gerloff, U., Tautz, D., Morin, P. A., Boesch, C., Fruth, B., Hohmann, G., Ryder, O. A. and Woodruff, D. S. (1999) Proc. Natl. Acad. Sci. U.S.A. **96**, 5077–5082
32 Morin, P. A., Moore, J. J., Chakraborty, R., Jin, L., Goodall, J. and Woodruff, D. S. (1994) Science **265**, 1193–1201
33 Hirsch, V. M., Campbell, B. J., Bailes, E., Goeken, R., Brown, C., Elkins, W. R., Axthelm, M., Murphy-Corb, M. and Sharp, P. M. (1999) J. Virol. **73**, 1036–1045
34 Beer, B. E., Bailes, E., Goeken, R., Dapolito, G., Coulibaly, C., Norley, S. G., Kurth, R., Gautier, J.- P., Gautier-Hion, A., Vallet, D., Sharp, P. M. and Hirsch, V. M. (1999) J. Virol. **73**, 7734–7744
35 Korber, B., Theiler, J. and Wolinsky, S. (1998) Science **280**, 1868–1871
36 Muller-Trutwin, M. C., Corbet, S., Dias Tavares, M., Herve, V. M. A., Nerrienet, E., Georges-Courbot. M.- C., Saurin, W., Sonigo, P. and Barre-Sinoussi, F. (1996) Virology **223**, 89–102
37 Uzzell, T. and Corbin, K. W. (1971) Science **172**, 1089–1096
38 Yang, Z. (1996) Trends Ecol. Evol. **11**, 367–372
39 Jin, L. and Nei, M. (1990) Mol. Biol. Evol. **7**, 82–102

40 Leitner, T., Kumar, S. and Albert, J. (1997) J. Virol. **71**, 4761–4770

41 Sharp, P. M., Robertson, D. L., Gao, F. and Hahn, B. H. (1994) AIDS **8**, S27–S42

42 Gao, F., Robertson, D. L., Morrison, S. G., Hui, H., Craig, S., Decker, J., Fultz, P. N., Girard, M., Shaw, G. M., Hahn, B. H. and Sharp, P. M. (1996) J. Virol. **70**, 7013–7029

43 Yang, Z. (1997) CABIOS **13**, 555–556

44 Eigen, M. and Nieselt-Struwe, K. (1990) AIDS **4**, S85–S93

45 Zhu, T., Korber, B. T., Nahmias, A. J., Hooper, E., Sharp, P. M. and Ho, D. D. (1998) Nature (London) **391**, 594–597

46 Sharp, P. M., Bailes, E., Robertson, D. L., Gao, F. and Hahn, B. H. (1999) Biol. Bull. **196**, 338–342

Received 9 September 1999