J Mol Evol (2000) 51:12–20
DOI: 10.1007/s002390010062

JOURNAL OF MOLECULAR EVOLUTION

Are RNA Viruses Adapting or Merely Changing?

Monica Sala, Simon Wain-Hobson

Unité de Rétrovirologie Moléculaire, Institut Pasteur, 28, rue du Dr. Roux, 75724 Paris Cedex 15, France

Received: 16 November 1999 / Accepted: 10 March 2000

Abstract. RNA viruses and retroviruses fix substitutions approximately 1 million-fold faster than their hosts. This diversification could represent an inevitable drift under purifying selection, the majority of substitutions being phenotypically neutral. The alternative is to suppose that most fixed mutations are beneficial to the virus, allowing it to keep ahead of the host and/or host population. Here, relative sequence diversification of different proteins encoded by viral genomes is found to be linear. The examples encompass a wide variety of retroviruses and RNA viruses. The smoothness of relative divergence spans quasispeciation following clonal infection, to variation among different isolates of the same virus, to viruses from different species or those associated with different diseases, indicating that the majority of fixed mutations likely reflects drift. This held for both mammalian and plant viruses, indicating that adaptive immunity doesn't necessarily shape the relative accumulation of amino acid substitutions. When compared to their hosts RNA viruses evolution appears conservative.

Key words: RNA virus — Retrovirus — Virus evolution — Sequence divergence — Genetic drift

Introduction

Viral replication is accompanied by destruction of huge numbers of progeny. This is particularly evident from recent work on the dynamics of three viruses: human immunodeficiency virus (HIV) (Ho et al. 1995; Wei et al. 1995) and human hepatitis B and C viruses (HBV and HCV) (Nowak et al. 1996; Zeuzem et al. 1998). Following therapeutic intervention, plasma viral loads were measured revealing massive destruction—between 10^8 to 10^{12} virions per day, or 50–90% of the total. Calculations related to the genesis of sequence diversity concur with these findings and indicate that the majority of infected cells are destroyed before they can give rise to progeny (Wain-Hobson 1993; Pelletier et al. 1995). In other words, only a small fraction of infected cells are productively infected. The obvious predators are immune responses. The other hurdle viruses have to overcome is transmission, which may represent a severe bottleneck—infection can be initiated by a single virion. The massive destruction and precariousness of transmission could well introduce a strong stochastic component into the evolution of their genomes.

In terms of genetic variation, a major distinction can be made between RNA and DNA viruses. In lieu of any nucleic acid proofreading mechanism, the mutation rates for RNA viruses and retroviruses are between 0.2–2 per genome per cycle (Drake 1991, 1993; Domingo and Holland 1997). This shows up in the amino acid fixation rates for RNA viral and retroviral proteins, which is of the order of 10^6 greater than those of their vertebrate, invertebrate, and plant hosts (Gojobori and Yokoyama 1985). By contrast, DNA viruses either encode proofreading enzymes (e.g., herpes and pox viruses) or are edited by the host replication machinery (e.g., papilloma viruses). Accordingly, the mutation rates for DNA viruses are 10^4–10^6-fold lower than their RNA counterparts (Domingo and Holland 1997). To what extent do

Correspondence to: S. Wain-Hobson

RNA viruses exploit genetic variation, if at all? Given such a superior fixation rate, are they adapting or merely changing?

A handful of studies have tracked viral sequence diversification over time, frequently revealing molecular clock–like behavior—that is, there was a linear increase in the number of fixed mutations over time, much like molecular clocks established for a variety of chromosomal genes. The examples included HIV, HBV, coronaviruses, and influenza A (Elena et al. 1992; Gojobori et al. 1990; Hayashida et al. 1985; Leitner et al. 1997; McGeoch et al. 1995; Plikat et al. 1997; Querat et al. 1990; Sanchez et al. 1992; Villaverde et al. 1991; Yang et al. 1995; Fitch et al. 1997). Such studies require clearly defined time points, which are not always available.

A way around this is to compare the extent of amino acid fixation from different viral proteins. Any two coding regions on the same genome will be replicating despite decimation by immune responses and the bottlenecking inherent to transmission. They may only become uncoupled via frequent recombination between divergent viruses. Hence, time is factored out. For a large and diverse selection of viruses it is shown here that amino acid substitutions for pairs of viral proteins are accumulated in a linear manner. This goes for vertebrate viruses as well as those of plants and even bacteriophages, indicating that this feature is not shaped by pressure from adaptive immune systems.

Materials and Methods

Sequences. Sequences were recovered from Los Alamos Web sites for HIV-1, HIV-2, and simian immunodeficiency virus (SIV) (http://hiv-web.lanl.gov) and papillomavirus (http://HPV-web.lanl.gov) as well as through GenBank and EMBL Bank for other viruses. All sequences used in this study as well as alignments are available at ftp://www.pasteur.fr/pub/retromol/Sala99 through anonymous login.

Relative Rates of Viral Gene Diversification. The study encompassed 85 different protein data sets, comprising 1174 sequences covering 15 groups of mammalian, plant, and bacterial viruses (see Table 1). The majority (92%) of sequences were taken from completely sequenced genomes. Data were used from some incomplete viral sequences, for example, the influenza hemagglutinins and some additional HIV-1, HIV-2, and SIV protein sequences. In general, only residues corresponding to a nonoverlapping portion of the open reading frame (orf) were used. For HBV the S and P orfs overlap extensively. This example was explored explicitly.

Sequences were aligned using ClustalW1.6. After gap stripping, divergence was calculated following weighting using Blosum matrices (Henikoff and Henikoff 1993). Viral genomes were paired arbitrarily, each genome being analyzed only once. Accordingly, all the data points were independent. Relative divergence was calculated in the following manner using HCV as an example: for a pair of HCV genomes the divergence among all homologous proteins (core, E1, etc.) were computed and plotted with respect to the divergence between NS5B proteins (viral polymerase) arbitrarily taken as reference. Comparisons of the relative divergence for pairs of proteins from other twinned HCV genomes provided additional points.

Results

HCV is a member of the Flaviviridae with a positive, single-stranded RNA genome averaging 9.4 kb. It includes two untranslated regions at the 5′ and 3′ ends, and a large orf encoding a ~3020-residue polyprotein, which is posttranslationally cleaved into structural (core, E1, E2) and nonstructural (NS2, NS3, NS4A, NS4B, NS5A, and NS5B) proteins. Thirty complete sequences were recovered from data banks, allowing 15 comparisons for any pair of mature proteins.

Figure 1A gives a series of comparisons for amino acid divergence among HCV E1, E2, NS2, NS3, NS4B, NS5A, and core proteins with respect to the NS5B protein (RNA polymerase) taken as reference. In all cases a linear relationship was noted, which proved to be statistically robust (Table 1). Furthermore the x and y intercepts were close to the origin. To show that linear relationships were not a chance characteristic of the way genomes were initially paired, the results of 15 alternative pairings for the NS2/NS5B pair are shown in Fig. 1B. As can be seen from Table 1, the difference among gradients was ≤12%. Finally, plotting dependent data, i.e., NS2/NS5B comparisons from all pairwise analyses yielded a linear relationship (Fig. 1C and Table 1) indistinguishable from the mean for all the 15 data sets shown in Fig. 1B.

No matter the protein, its relative divergence with respect to the NS5B protein was linear out to Blosum distances of 0.1–0.5 and 0.3 for NS5B. Blosum corrections were introduced as protein sequence comparisons are more informative when weights are used based on genetic and structural biases for amino acid replacements (Henikoff and Henikoff 1993). To explore the impact of Blosum corrections on the relative divergence of viral proteins, pairwise comparisons were made for the HCV NS5A/NS5B and E2/NS5B without any correction. As can be seen from Table 1, the linear relationships, particularly the gradients, were hardly affected. Hence, relative divergence seems to be a smooth process, the rate being an intrinsic property of the proteins.

The absence of points far from the linear regression substantiates the assumption that recombination between highly divergent genomes is a rare phenomenon. This doesn’t deny recombination, merely that the frequency of homologous recombination is probably a function of sequence divergence. Hence the closer the sequences the more probable the event. Such examples would not result in a large deviation from a linear regression.

The approach was extended to a number of different viruses for which there were a substantial number of complete sequences. Not surprisingly, HIV-1, HIV-2, and SIV provide a particularly voluminous data base. Using p66 reverse transcriptase (RT) as reference, significant linear relationships were obtained for comparisons with the p17 Gag, p24 Gag, integrase, gp41, gp120, Vif, and Nef proteins (Table 1). Furthermore, linear re-

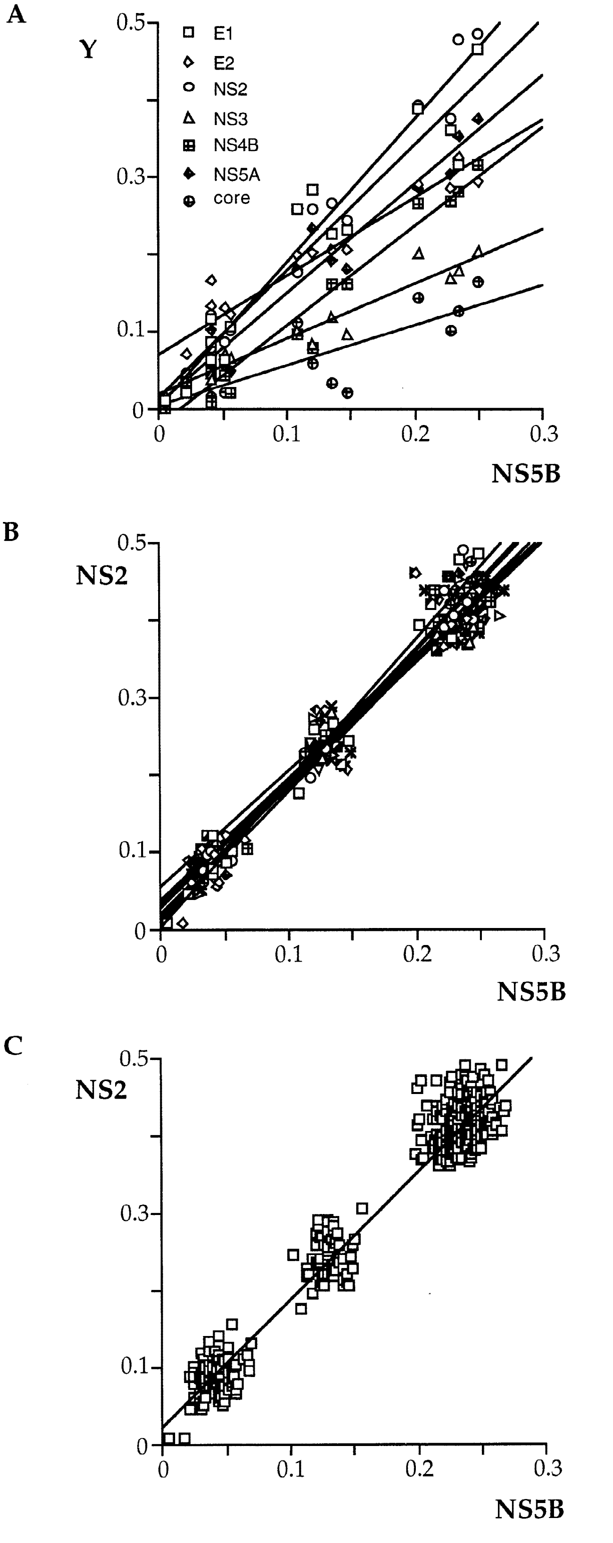

Fig. 1. Graphical representation of paired divergence for orthologous proteins taken from complete HCV genomes. X and Y values correspond to Blosum-corrected fractional divergence. The straight lines were obtained by linear regression analysis. Their characteristics are given in Table 1. **A** Y = different proteins, X = NS5B protein. **B** 15 independent NS2/NS5B data sets superposed. Each graph is indicated as NS2.n, where n = 1 to 15 (Table 1). **C** NS2/NS5B dependent data set.

gression passed close to the origin except for a plot of Nef versus RT. Despite this, the correlation coefficient was highly significant. As for HCV, the relative rates of sequence divergence (y/x) varied by little more than a factor of two. The reason why the hypervariable gp120 protein shows a relatively low degree of change with respect to RT is that gap stripping effectively eliminates the hypervariable regions. Consequently, the analysis pertains to the conserved regions of gp120. All other alignments were only marginally influenced by gap stripping. Using larger data sets from partially sequenced HIV/SIV genomes yielded linear relationships for p17 Gag/p24 Gag, integrase/RT, and gp41/gp120 (Table 1).

This HIV-1, HIV-2, and SIV data set is particularly interesting as it covers the earliest phase of genetic diversification (intrapatient variation, generally <10% at the nucleic acid level), continuing smoothly to cover interclade, intertype, and finally interspecies comparisons. The same forces are apparently uppermost during all stages of diversification. Yet this in spite of different environments—that of an individual's immune system, different immune systems stigmatized by highly polymorphic HLA, and finally differences between humans, chimpanzees, mandrills, and African green monkeys accumulated over 30 million years. It is remarkable that such very different proteins as gp120 and the gp41 ectodomain (surface glycoproteins), p17 Gag and p24 Gag (structural proteins), RT and integrase (enzymes), and Nef and Vif (cytoplasmic proteins), all yielded linear relationships (Table 1) as though the fixation amino acid substitutions were an intrinsic property of each protein.

Highly significant linear regression relationships were also in evidence for another set of retroviruses, human type 1 and 2 and simian T-cell leukemia viruses (HTLV-1, HTLV-2, and STLV-1). Despite the fact that immunodeficiency viruses and HTLVs infect the same host (human and simian) and the same cell (CD4+ lymphocytes, HTLV-1 and STLV-1), they differ markedly in their viral cycles. HTLVs replicate essentially through clonal expansion of the infected cell (Wattel et al. 1995; Cimarelli et al. 1996), whereas immunodeficiency viruses achieve high proviral loads through hundreds of rounds of reverse transcription per year. As a consequence, HIV fixes substitutions approximately a thousand fold faster than HTLV (Mahieux et al. 1997). Despite this, relative divergence of proteins from both groups of viruses was linear.

When extended to other viruses, the approach yielded linear relationships in all cases. Whether the data was for influenza A viruses, hepatitis B, E, or G viruses, rhinoviruses, enteroviruses, or flaviviruses (other than HCV), linear relationships were found (Table 1). The influenza A hemagglutinin surface and transmembrane protein

Table 1. Pairwise amino acid sequence divergence for a wide collection of viruses

Virus Paired proteins (y/x)	n	Linear relationship	Correlation coefficient: r
HCV			
E1/NS5B	15	y = 1.64x + 0.02	0.96*
E2/NS5B	15	y = 1.01x + 0.07	0.96*
NS2/NS5B	15	y = 1.87x + 0.00	0.99*
NS3/NS5B	15	y = 0.71x + 0.02	0.96*
NS4B/NS5B	15	y = 1.28x − 0.02	0.98*
NS5A/NS5B	15	y = 1.42x + 0.01	0.98*
core/NS5B	15	y = 0.52x + 0.00	0.84*
NS5A/NS5B Blosum matrix	15	y = 1.42x + 0.01	0.98*
NS5A/NS5B Identity matrix	15	y = 1.46x + 0.01	0.98*
E2/NS5B Blosum matrix	15	y = 1.01x + 0.07	0.96*
E2/NS5B Identity matrix	15	y = 1.02x + 0.07	0.97*
NS2/NS5B (dd)	435	y = 1.65x + 0.02	0.98*
NS2/NS5B.1	15	y = 1.87x + 0.00	0.99*
NS2/NS5B.2	15	y = 1.77x + 0.01	0.97*
NS2/NS5B.3	15	y = 1.77x + 0.01	0.99*
NS2/NS5B.4	15	y = 1.63x + 0.02	0.98*
NS2/NS5B.5	15	y = 1.60x + 0.03	0.97*
NS2/NS5B.6	15	y = 1.64x + 0.03	0.99*
NS2/NS5B.7	15	y = 1.74x + 0.01	0.99*
NS2/NS5B.8	15	y = 1.56x + 0.04	0.98*
NS2/NS5B.9	15	y = 1.50x + 0.05	0.97*
NS2/NS5B.10	15	y = 1.59x + 0.03	0.98*
NS2/NS5B.11	15	y = 1.58x + 0.03	0.98*
NS2/NS5B.12	15	y = 1.74x + 0.01	0.97*
NS2/NS5B.13	15	y = 1.64x + 0.03	0.97*
NS2/NS5B.14	15	y = 1.58x + 0.04	0.96*
NS2/NS5B.15	15	y = 1.70x + 0.01	0.96*
Means		1.66 0.02	0.98*
HIV-1, -2, SIV			
p17/RT	22	y = 1.15x − 0.02	0.98*
p24/RT	22	y = 0.91x + 0.03	0.96*
integrase/RT	22	y = 0.94x − 0.01	0.99*
gp41/RT	22	y = 1.16x + 0.01	0.96*
gp120/RT	22	y = 1.50x + 0.02	0.96*
vif/RT	22	y = 1.49x + 0.01	0.97*
nef/RT	22	y = 1.19x + 0.12	0.94*
p17/p24	53	y = 1.01x + 0.09	0.93*
integrase/RT	24	y = 0.95x − 0.01	0.96*
gp41/gp120	63	y = 0.79x − 0.02	0.97*
nef/vif	23	y = 0.71x + 0.12	0.88*
HTLV-1, -2, STLV			
p24/RT	7	y = 0.35x + 0.02	0.88*
p19/RT	7	y = 1.09x + 0.00	0.99*
integrase/RT	7	y = 0.96x + 0.02	0.98*
gp21/RT	7	y = 0.35x + 0.01	0.96*
gp46/RT	7	y = 0.83x + 0.01	1
HBV			
pol/preS1-preS2-HBsAg	17	y = 1.07x + 0.00	0.96*
pol overl/preS1-preS2-HBsAg	17	y = 0.85x + 0.00	0.97*
pol overl/HBsAg	17	y = 1.11x + 0.00	0.94*
HEV			
M/polymerase	5	y = 0.78x + 0.00	0.96^{+}
Y/polymerase	5	y = 0.51x + 0.02	0.45^{-}
Pr/polymerase	5	y = 2.56x − 0.01	0.99*
P/polymerase	5	y = 4.40x + 0.01	0.98*
X/polymerase	5	y = 1.75x − 0.01	1
H/polymerase	5	y = 0.91x + 0.00	0.98*
ORF2/polymerase	5	y = 0.48x + 0.01	0.99*
Influenza A viruses			
HA1/HA2	47	y = 1.32x + 0.03	0.97*
HGV			
E1/NS5A	5	y = 1.06x + 0.00	0.78$^{\$}$
E2/NS5A	5	y = 1.36x + 0.02	0.68^{-}
NS2/NS5A	5	y = 1.68x + 0.00	0.94$^{\#}$
NS3/NS5A	5	y = 0.57x + 0.00	0.96^{+}
NS4/NS5A	5	y = 0.85x + 0.00	0.69^{-}
NS5B/NS5A	5	y = 0.84x + 0.00	0.81$^{\$}$
Rhinoviruses			
3C/3D	5	y = 1.05x − 0.01	0.98*
P2A/3D	5	y = 1.20x − 0.04	0.91″
P2C/3D	5	y = 1.14x − 0.02	0.99*
VP1/3D	5	y = 1.27x + 0.02	0.99*
VP2/3D	5	y = 0.77x + 0.03	0.97^{+}
VP3/3D	5	y = 0.96x + 0.02	0.95^{+}
Enteroviruses			
3C/3D	14	y = 1.39x + 0.00	0.99*
P2A/3D	14	y = 1.22x + 0.05	0.93*
P2B/3D	14	y = 1.43x + 0.03	0.97*
P2C/3D	14	y = 1.28x + 0.00	0.98*
VP1/3D	14	y = 1.44x + 0.16	0.95*
VP2/3D	14	y = 1.03x + 0.12	0.94*
VP3/3D	14	y = 1.14x + 0.12	0.92*
Flaviviruses			
C/NS5	20	y = 1.77x + 0.00	0.98*
prM/NS5	20	y = 1.60x + 0.00	0.99*
E/NS5	20	y = 1.42x + 0.00	0.99*
NS1/NS5	20	y = 1.40x + 0.00	1
NS2A/NS5	20	y = 1.89x + 0.03	0.98*
NS2B/NS5	20	y = 1.82x + 0.01	0.99*
NS3/NS5	20	y = 1.22x − 0.01	0.99*
NS4A/NS5	20	y = 1.64x + 0.00	1
NS4B/NS5	20	y = 1.97x + 0.00	0.95*
Potexviruses (plants)			
capsid/pol	6	y = 1.40x + 0.03	0.98*
ORF2/pol	6	y = 1.28x + 0.02	0.99*
Potyviruses (plants)			
P1/CI	21	y = 1.42x + 0.15	0.94*
HC-Pro/CI	21	y = 1.26x − 0.01	0.99*
P3/CI	21	y = 1.53x + 0.06	0.97*
NIa/CI	21	y = 1.13x + 0.00	0.99*
NIb/CI	21	y = 0.87x + 0.01	0.99*
CP/CI	21	y = 0.92x − 0.01	0.98*
Papillomaviruses			
L2/L1 Blosum matrix	25	y = 1.29x − 0.02	0.96*
L2/L1 Identity matrix	25	y = 1.28x − 0.01	0.96*
E1/L1 Blosum matrix	25	y = 1.02x + 0.08	0.90*
E1/L1 Identity matrix	25	y = 1.01x + 0.08	0.89*
Geminiviruses (plants)			
V2(AR1)/C1(AL1)	17	y = 1.21x − 0.06	0.80*
Inoviridae (bacteria)			
protein I/II (dd)	15	y = 1.06x + 0.03	1
protein III/II (dd)	15	y = 1.02x + 0.11	0.91*
protein IV/II (dd)	15	y = 1.02x + 0.05	0.99*

HIV/SIV: human and simian immunodeficiency viruses; HBV, HCV, HEV, HGV: human hepatitis B, C, E, and G viruses, respectively; HTLV/STLV: human and simian T-cell leukemia viruses. The protein comparisons made for paired genomes are listed as y and x and identify each in the linear regression analyses. The numbers of paired genomes per virus group is n. The significance of the correlation coefficients is as follows: * $p < 0.001$, $^{+}$ $p < 0.005$, $^{\#}$ $p < 0.01$, ″ $p < 0.02$, $^{\$}$ $p < 0.1$, and $^{-}$ $p \geq 0.1$. For HCV there were 15 arbitrarily chosen pairs of genomes, which yielded as many independent data points (NS2/NS5B.n, where n = 1–15). Occasionally all pairwise comparisons were made resulting in dependent data (dd); see also the inoviridae. For HBV the comparisons are for the entire pol (P) and the PreS1-PreS2-HBsAg orf, uniquely the overlapping region of these two orfs and the P region overlapping only the HBsAg coding region, respectively.

(HA1/HA2) pair is salutary as it is considered the paradigm for positive selection (Fitch et al. 1997). Others have long noted a molecular clock behavior for the hemagglutinin and some structural proteins as well as the accumulation of synonymous substitutions (Gojobori et al. 1990; Fitch et al. 1997).

All the above viruses infect mammals endowed with powerful adaptive and innate immune systems. Plants and insects are hosts to a wide variety of viruses and defend themselves uniquely via a multitude of innate defense systems markedly different to those of mammals (Brey and Hultmark 1998). Although there are insufficient data to allow a comparable analysis of insect viruses, there are large numbers of complete plant virus sequences available. When applied to proteins encoded by either potexviruses or potyviruses (positive stranded RNA genomes), relative divergence was always linear with intercepts close to the origin. These examples show that the adaptive immune system does not a priori underlie the linear relationships. Again, the degree of variation seems to be more a feature of the particular protein, rather than the host species.

Among all the different protein comparisons, very few failed to yield significant linear relationships (e.g., HGV and HEV data sets, Table 1). These failures had two common features: a small sample size ($n = 5$) and low divergence (Blosum distances <0.1). It would seem to be the conjunction of the two features, for all rhinovirus ($n = 5$) comparisons were significant (Blosum distances <0.7) as were the potexvirus ($n = 6$) comparisons (Blosum distances <0.8). Consequently, it is possible that as more and more HEV and HGV genomes are sequenced comparisons of the divergence among pairs of proteins will prove to be significant in a linear regression analysis.

As DNA replication is subject to proofreading, DNA viruses fix substitutions at a much reduced rate compared to their RNA counterparts (Drake 1991, 1993; Domingo and Holland 1997). This apart, might not the relative accumulation of sequence diversity among DNA viral proteins parallel the findings for RNA viruses? When applied to papillomavirus (double-stranded DNA viruses) linear relationships were to be had (Fig. 2, Table 1). Note the large Blosum distances and a particularly diverse collection spanning oncogenic or benign human, chimpanzee, bovine, canine, deer, and elk viruses. The same was true for a pair of proteins encoded by the double-stranded DNA geminiviruses of plants, again indicating that linearity has nothing to do per se with the adaptive immunity. A final example is provided by the inoviruses, single-stranded DNA bacteriophages of the fd group, which includes M13. Comparisons of their I, II, III, and IV protein sequences yielded linear relationships (Table 1). And of course, bacterial defenses against infecting viruses are markedly different from both mammalian and plant immune responses.

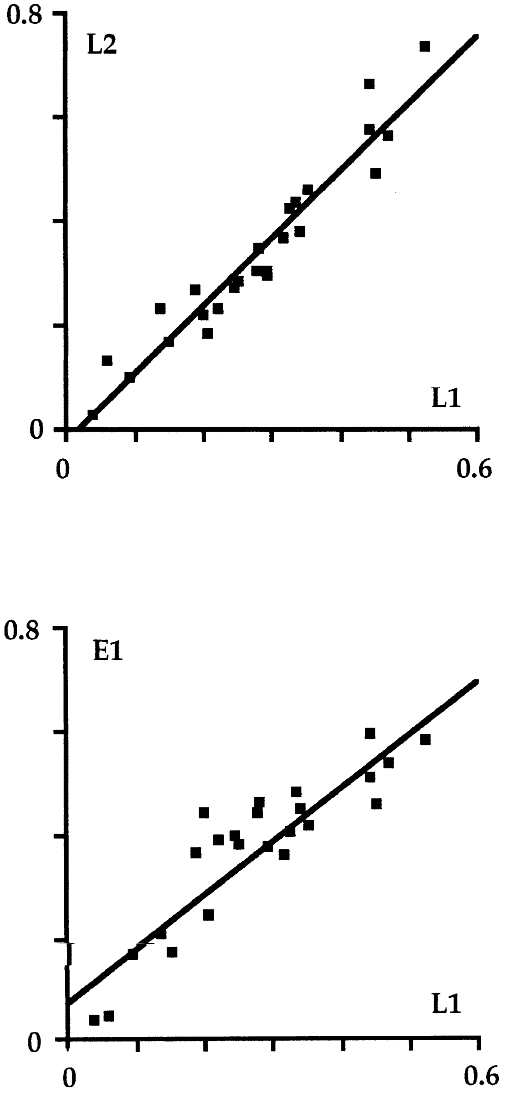

Fig. 2. Graphical representation of paired divergence for orthologous proteins taken from complete papillomavirus genomes encompassing 45 human, 3 bovine, and 1 dog, rabbit, and deer viruses. Units correspond to Blosum-corrected fractional divergence. The straight lines were obtained by linear regression analysis (Table 1). Given the extent of papillomavirus sequence variation, the analysis was also performed using uncorrected amino acid sequence differences. The results for the L2/L1 and E1/L1 pairs were almost the same (Table 1) as those using Blosum distances, again indicating that the Blosum correction was not itself responsible for the phenomenon.

Discussion

Whether the comparisons were between capsid proteins versus enzymes, secretary versus cytoplasmic molecules, or proteins of very different three-dimensional structure, significant linear relationships were obtained in all cases. This pertains to proteins encoded by both RNA and DNA viruses replicating in mammals, plants, and bacteria. Furthermore, the intercepts were very close to the origin, indicating a smooth process spanning quasispeciation following clonal infection, to variation among different isolates of the same virus, to viruses from different species or those associated with different diseases. Although linear relationships were maintained out to large Blosum distances, they were not dependent on this correction (e.g., HCV and papillomaviruses; Table 1 and unpublished data).

What is the significance of the linear relationships? The gradients correspond to ratios of amino acid fixation for pairs of viral proteins and are therefore scalar quantities. However, they are indistinguishable from the rate of amino acid fixation for protein A divided by that for protein B. Formally, the rate of fixation need not be constant, i.e., clocklike, so long as concurrent rate changes showed up among all viral proteins studied. However, molecular clocks for viral proteins have been observed over small degrees of divergence (Elena et al. 1992; Gojobori et al. 1990; Hayashida et al. 1985; Leitner et al. 1997; McGeoch et al. 1995; Plikat et al. 1997; Querat et al. 1990; Sanchez et al. 1992; Villaverde et al. 1991; Yang et al. 1995), certainly for a number of viruses analyzed here (e.g., influenza A hemagglutinin (HA), HIV Nef, HBV HBsAg). Hence it may be concluded that for small Blosum distances, out to 0.2–0.3, the linearity of relative sequence divergence is reflecting molecular clock behavior.

What may be said for the fixation at larger distances when multiple substitutions must be occurring? (1) For a data set spanning 64 orthologous sets of chromosomal proteins, a molecular clock behavior was evident after applying the Blosum correction (Doolittle et al. 1996; Feng et al. 1997). Therefore plotting Blosum-corrected distances for paired proteins would be expected to yield a straight line. (2) For the same data set (Doolittle et al. 1996; Feng et al. 1997) it can be shown that the uncorrected distances vary with time according to the relationship $\Delta A = at^m$, where ΔA is the sequence divergence among evolutionary related members of the set of proteins A, and a and m are constants. Similarly for protein B the relationship would be $\Delta B = bt^n$, where b and n are constants. Obviously, since the maximum values of ΔA and ΔB tend to 1, m, n < 1. Comparing divergence among two proteins, say, A and B, encoded by the same genome yields $\Delta A/\Delta B = a/bt^{(m-n)}$. The format of the presentations in Figs. 1 and 2 is that of ΔA versus ΔB. A linear relationship is only possible when m = n, while the gradient equals the ratio of two constants, a and b. Rewriting the equations for a set of proteins one obtains $\Delta A = at^m$, $\Delta B = bt^m \ldots \Delta Z = zt^m$. Hence, differences in the rate of fixation of amino acid substitutions for different proteins are directly related to a constant (a, b . . . z), intrinsic to each protein.

A simple hypothesis to explain the smoothness of protein sequence diversification over a wide variety of differing hosts or niches is that the *majority* of fixed amino acid substitutions reflect drift under purifying selection, the rate of accumulation being intrinsic to each protein. This is not to say that positive selection is inoperative—the case of influenza A HA1 being a case in point—merely that the *majority* of fixed substitutions are essentially neutral (Haydon et al. 1998), so much so that it does not strongly distort the data from a linear relationship expected for genetic drift.

By contrast, it could be postulated that positive selection is continually operative on all proteins. The obvious selection pressure on all mammalian viral proteins is T-cell immunity, which may recognize processed peptides from all viral proteins. However, linearity was noted for the relative divergence of plant viruses and bacteriophages, indicating that viruses such as HIV, HCV, and influenza A are not necessarily mutating to outwit their host's immune system. Or in terms of the above language, it is not axiomatic that the majority of amino acid changes reflects escape from host adaptive immune systems. In fact, an examination of the literature reveals very few unambiguous examples of escape from adaptive immune systems by ongoing variation within the host (Wain-Hobson 1996; Borrow and Shaw 1998). Adaptive immunity is intimately associated with memory, which allows rapid secondary responses and protection of offspring by transfer of antibodies via the placenta and milk (Zinkernagel 1996; Mason 1998).

Pursuing the argument in favor of continuous positive selection, what other mechanism(s) could be envisaged that are compatible with the smoothness of relative divergence in a protein specific manner? Fitness selection is an alternative, but presents a conundrum. For example, the data in Fig. 1 were all for the human hepatotropic virus, HCV. Being able to exclude T-cell immunity, it becomes reasonable to assume that the niche is fairly constant. Yet to conclude fitness selection is to assume that HCV is not adapted to its niche despite a 10^6-fold increased mutation rate over host cell replication. Furthermore, given such an advantage, continual fitness selection, unrelated to T-cell immunity, should be measurable over a decade or so, then one would expect a steady increase in virulence. A similar argument could be made for HBV, HEV, or HGV. There is no evidence that HIV-1 is increasing in virulence. In contrast, the HIV/SIV, rhino-, entero-, flavi-, influenza A, potex-, poty-, papilloma-, and geminiviruses data sets included viruses from different hosts and in some cases viruses of different tropism (Table 1, Fig. 2, ftp site). It is safe to conclude that some degree of adaptation must have occurred. Despite this, linear relationships were found for all cases.

The data support the hypothesis that the *majority* of fixed substitutions are phenotypically neutral. The number of substitutions allowing adaptation cannot be so great, otherwise the lineage would become extinct. Since viral populations sizes in vivo are rarely $>10^{12}$ at any moment, the number of mutations necessary to allow adaptation cannot be more than three per genome, given

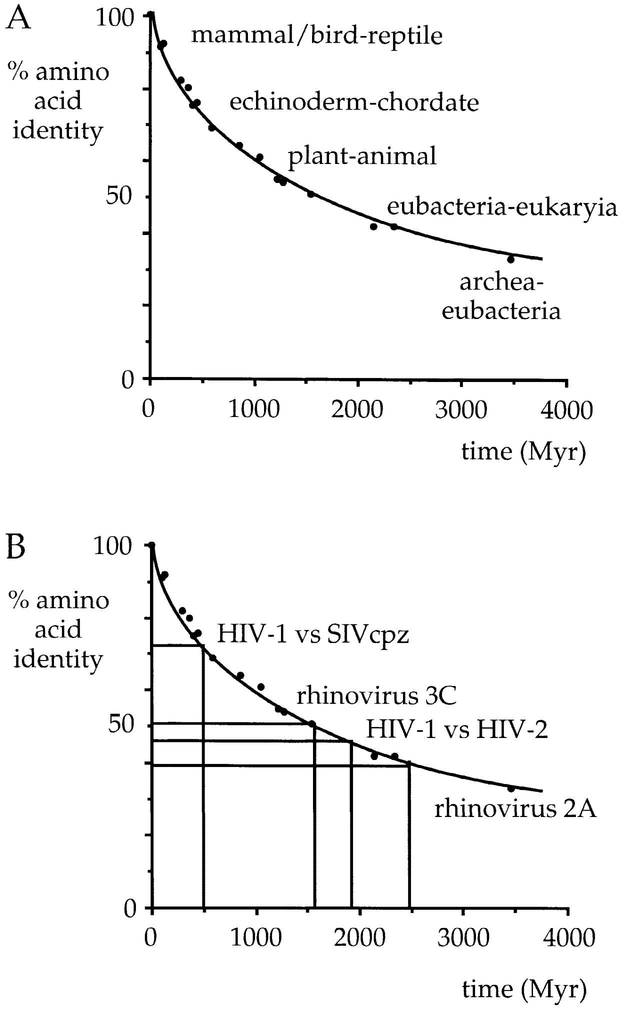

Fig. 3. **A** Mean percent sequence divergence for an ensemble of 64 sets of proteins over geological time. The time of major divisions in evolution are noted. Adapted from (Doolittle et al. 1996; Feng et al. 1997). **B** The percent sequence divergence for pairs of viral proteases are projected onto the geological time scale.

viral mutation frequencies of 10^{-4} to 10^{-5} per base per cycle and a genome length of ~10^4 bases. Of course, if the effective population size was much lower than the census population, as some studies suggest (Leigh Brown 1997), the requirements become more stringent. Numerous studies of protein mutagenesis have shown that the fraction of viable mutants is surprisingly large while only a handful of substitutions are sufficient to confer enhanced thermal stability or even novel catalytic activities (Klein et al. 1997; Olins et al. 1995; Rennell et al. 1991; Martinez et al. 1996; Kucher and Arnold 1997; Quéméneur et al. 1998). The acquisition of drug resistance is the affair of a handful of residues, which is readily achieved by an RNA virus or retrovirus. The number is a trivial fraction of the total possible or even the total that can be surmised from known lineages (Maynard Smith 1970).

The present data indicate that viral proteins accumulate substitutions in a concerted manner. The major variable appears to be the nature of the protein as shown by differing slopes (Figs. 1, 2, Table 1). The analogy with the differential accumulation of substitutions among fibronectins, globins, and cytochromes is striking. These are the classic examples used to illustrate regular molecular clock–like behavior (Zuckerkandl 1997; Wilson et al. 1977, 1987; Kimura 1983). It is obvious that if the divergence among pairs of α-globin sequences are compared to that for pairs of cytochrome c from the same genomes, then a straight line would result. The interest of RNA viruses and retroviruses is their much enhanced mutation and fixation rates.

It is possible to compare the evolution of viral and host proteins in view of the 10^6-fold difference in mutation and fixation rates? In a study of sequence variation for 64 groups of enzymes over geological time scales (Fig. 3A) different enzymes fixed substitutions at rates that varied by little more than a factor of 2–3 (Doolittle et al. 1996; Feng et al. 1997). The parallel with viral proteins is striking: the range of relative rates was rarely more than a factor of 2–3-fold (gradients in Table 1), which indicates that there is little qualitative difference in the way viral proteins fix substitutions. The comparison of sequence divergence among orthologous viral and cellular proteins suggests a means to compare the two.

RNA viruses encode a number of proteases that are orthologous to well-known serine, cysteine, or aspartic acid proteases (Babé and Craik 1997). For example, all retroviruses encode an aspartic protease whose Cα coordinates overlap remarkably those of pepsin (Wlodawer and Erickson 1993). Equally the picornaviruses and flaviviruses encode proteases orthologous to serine and cysteine proteases (Ryan and Flint 1997; Ryan et al. 1998). Among the 64 groups of enzymes analyzed was trypsin, a serine protease. Albeit absent from the original study, it is a reasonable assumption that the fixation of replacement substitutions for the chromosomal cysteine or aspartic acid proteases falls within the range described. Superimposing rhinovirus 2A or 3C proteases sequence divergence onto this figure is tantamount to >700 million years of chromosomal protein evolution (Fig. 3B). Sequence variation within the HIV-1 aspartic protease alone is equivalent to >300 million years. Other examples reinforce these observations. In fact some 10 years ago it was noted that sequence variation among retroviral aspartic acid proteases was comparable to billions of years of DNA-based chromosomal evolution (Doolittle et al. 1989). Figure 3B extends this thesis.

Yet there exists a major difference. The enzyme sequences were taken from a highly diverse group of organisms, ranging from bacteria to plants and humans. The viral counterparts illustrated occupy, so far as is known, a single niche. All HIV-1 genomes are isogenic. The same is true for many other viruses listed in Table 1. By contrast autonomous microbial, plant, and human genomes differ hugely in organization (e.g., number of chromosomes, gene inventories and families, synteny) despite a 4–6-log handicap in terms of point mutation rate.

Why is the evolution of RNA viral genome organization so conservative with respect to that of their cellular hosts? This probably stems from the size limit inherent to RNA viral genomes. Due to the lack of proofreading the range is generally 5–15 kb, with the coronaviruses providing an upper limit at around 30 kb. Generally 50–100% of a RNA viral genome is taken up with coding for structural proteins, enzymes, and proteins involved in gene regulation. There is little room for additional coding material or gene duplication with subsequent diversification. DNA viral replication involves proofreading either by the host cell machinery or by viral encoded enzymes. DNA viruses genomes can be as small as the ~2-kb circuloviruses, extending up to the pox and herpes viruses (70–240 kb), which have picked up host cell genes (Smith et al. 1997). One of the largest viral genomes, that of bacteriophage G (~670 kb), is larger than that of the smallest cellular microbe, *Mycoplasma genitalium* (580 kb). Among many other things, the explosion in bacterial genomics is revealing large numbers of multigene families, insertion elements, phage remnants, and pathogenicity islands (Hacker et al. 1997). By not encoding proofreading enzymes, RNA viruses cannot explore large genome configurations allowing gene duplication, domain shuffling, and gene capture.

In conclusion, it appears that RNA viruses change more than they adapt.

Acknowledgments. This work was supported by grants from the Institut Pasteur and l'Agence Nationale pour la Recherche sur le SIDA (ANRS).

References

Babé LM, Craik CS (1997) Viral proteases: evolution of diverse structural motifs to optimize function. Cell 91:427–430

Borrow P, Shaw GM (1998) Cytotoxic T-lymphocyte escape viral variants: how important are they in viral evasion of immune clearance in vivo? Immunol Rev 164:37–51

Brey PT, Hultmark D (1998) Molecular mechanisms of immune responses in insects. Chapman & Hall, London

Cimarelli A, Duclos CA, Gessain A, Casoli C, Bertazzoni U (1996) Clonal expansion of human T-cell leukemia virus type II in patients with high proviral load. Virology 223:362–364

Domingo E, Holland JJ (1997) RNA viral mutations and fitness for survival. Annu Rev Microbiol 51:151–178

Doolittle RF, Feng DF, Johnson MS, McClure MA (1989) Origins and evolutionary relationships of retroviruses. Quart Rev Biol 64:1–30

Doolittle RF, Feng D-F, Tsang S, Cho G, Little E (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–477

Drake JW (1991) A constant rate of spontaneous mutation in DNA-based microbes. Proc Natl Acad Sci USA 88:7160–7164

Drake JW (1993) Rates of spontaneous mutations among RNA viruses. Proc Natl Acad Sci USA 90:4171–4175

Elena SF, Gonzalez-Candelas F, Moya A (1992) Does the VP1 gene of foot-and-mouth disease virus behave as a molecular clock? J Mol Evol 35:223–229

Feng D-F, Cho G, Doolittle RF (1997) Determining divergence times with a protein clock: update and reevaluation. Proc Natl Acad Sci USA 94:13028–13033

Fitch WM, Bush RM, Bender CA, Cox NJ (1997) Long term trends in the evolution of H(3) HA1 human influenza type A. Proc Natl Acad Sci USA 94:7712–7718

Gojobori T, Yokoyama S (1985) Rates of evolution of the retroviral oncogene of Moloney murine sarcoma virus and of its cellular homologues. Proc Natl Acad Sci USA 82:4198–4201

Gojobori T, Moriyama EN, Kimura M (1990) Molecular clock of viral evolution, and the neutral theory. Proc Natl Acad Sci USA 87:10015–10018

Hacker J, Blum-Oehler G, Muhldorfer I, Tschape H (1997) Pathogenicity islands of virulent bacteria: structure, function and impact on microbial evolution. Mol Microbiol 23:1089–1097

Haydon D, Lea S, Fry L, Knowles N, Samuel AR, Stuart D, Woolhouse ME (1998) Characterizing sequence variation in the VP1 capsid proteins of foot and mouth disease virus (serotype 0) with respect to virion structure. J Mol Evol 46:465–475

Hayashida H, Toh H, Kikuno R, Miyata T (1985) Evolution of influenza virus genes. Mol Biol Evol 2:289–303

Henikoff S, Henikoff JG (1993) Performance evaluation of amino acid substitution matrices. Proteins 17:49–61

Ho DD, Neumann AU, Perelson AS, Chen W, Leonard JM, Markowitz M (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature 373:123–126

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Klein BK, Feng Y, McWherter CA, Hood WF, Paik K, McKearn JP (1997) The receptor binding site of human interleukin-3 defined by mutagenesis and molecular modeling. J Biol Chem 272:22630–22641

Kucher O, Arnold FH (1997) Directed evolution of enzyme catalysts. Trends Biotechnol 15:523–530

Leigh Brown AJ (1997) Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci USA 94:1862–1865

Leitner T, Kumar S, Albert J (1997) Tempo and mode of nucleotide substitutions in gag and env gene fragments in human immunodeficiency virus type 1 populations with a known transmission history. J Virol 71:4761–4770

Mahieux R, Ibrahim F, Mauclere P, Herve V, Michel P, Tekaia F, Chappey C, Garin B, Van Der Ryst E, Guillemain B, Ledru E, Delaporte E, de The G, Gessain A (1997) Molecular epidemiology of 58 new African human T-cell leukemia virus type 1 (HTLV-1) strains: identification of a new and distinct HTLV-1 molecular subtype in Central Africa and in Pygmies. J Virol 71:1317–1333

Martinez MA, Pezo V, Marliere P, Wain-Hobson S (1996) Exploring the functional robustness of an enzyme by in vitro evolution. EMBO J 15:1203–1210

Mason D (1998) A very high level of crossreactivity is an essential feature of the T-cell receptor. Immunol Today 19:395–404

Maynard Smith J (1970) Natural selection and the concept of a protein space. Nature 225:563–564

McGeoch DJ, Cook S, Dolan A, Jamieson FE, Telford EA (1995) Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. J Mol Biol 247:443–458

Nowak MA, Bonhoeffer S, Hill AM, Boehme R, Thomas HC, McDade H (1996) Viral dynamics in hepatitis B virus infection. Proc Natl Acad Sci USA 93:4398–4402

Olins PO, Bauer SC, Braford-Goldberg S, Sterbenz K, Polazzi JO, Caparon MH, Klein BK, Easton AM, Paik K, Klover JA, Thiele BR, McKearn JP (1995) Saturation mutagenesis of human interleukin-3. J Biol Chem 270:23754–23760

Pelletier E, Saurin W, Cheynier R, Letvin NL, Wain-Hobson S (1995)

The tempo and mode of SIV quasispecies development in vivo calls for massive viral replication and clearance. Virology 208:644–652

Plikat U, Nieselt-Struwe K, Meyerhans A (1997) Genetic drift can dominate short-term human immunodeficiency virus type 1 nef quasispecies evolution in vivo. J Virol 71:4233–4240

Quéméneur E, Moutiez M, Charbonnier J-B, Ménez A (1998) Engineering cyclophilin into a proline-specific endopeptidase. Nature 391:301–304

Querat G, Audoly G, Sonigo P, Vigne R (1990) Nucleotide sequence analysis of SA-OMVV, a visna-related ovine lentivirus: phylogenetic history of lentiviruses. Virology 175:434–447

Rennell D, Bouvier SE, Hardy LW, Poteete AR (1991) Systematic mutation of bacteriophage T4 lysozyme. J Mol Biol 222:67–88

Ryan MD, Flint M (1997) Virus-encoded proteinases of the picornavirus super-group. J Gen Virol 78:699–723

Ryan MD, Monaghan S, Flint M (1998) Virus-encoded proteinases of the Flaviviridae. J Gen Virol 79:947–959

Sanchez CM, Gebauer F, Sune C, Mendez A, Dopazo J, Enjuanes L (1992) Genetic evolution and tropism of transmissible gastroenteritis coronaviruses. Virology 190:92–105

Smith GL, Symons JA, Khanna A, Vanderplasschen A, Alcami A (1997) Vaccinia virus immune evasion. Immunol Rev 159:137–154

Villaverde A, Martinez MA, Sobrino F, Dopazo J, Moya A, Domingo E (1991) Fixation of mutations at the *VP1* gene of foot-and-mouth disease virus. Can quasispecies define a transient molecular clock? Gene 103:147–153

Wain-Hobson S (1993) Viral burden in AIDS. Nature 366:22

Wain-Hobson S (1996) Running the gamut of retroviral variation. Trends Microbiol 4:135–141

Wei X, Ghosh SK, Taylor ME, Johnson VA, Emini EA, Deutsch P, Lifson JD, Bonhoeffer S, Nowak MA, Hahn BH, Saag MS, Shaw GM (1995) Viral dynamics in human immunodeficiency virus type 1 infection. Nature 373:117–122

Wilson AC, Carlson SS, White TJ (1977) Biochemical evolution. Ann Rev Biochem 46:573–639

Wilson AC, Ochman H, Prager EM (1987) Molecular time scale for evolution. Trends Genet 3:241–247

Wlodawer A, Erickson JW (1993) Structure-based inhibitors of HIV-1 protease. Ann Rev Biochem 62:543–585

Yang Z, Lauder IJ, Lin HJ (1995) Molecular evolution of the hepatitis B virus genome. J Mol Evol 41:587–596

Zeuzem S, Lee JH, Franke A, Ruster B, Prummer O, Herrmann G, Roth WK (1998) Quantification of the initial decline of serum hepatitis C virus RNA and response to interferon alfa. Hepatology 27:1149–1156

Zinkernagel RM (1996) Immunology taught by viruses. Science 271: 173–178

Zuckerkandl E (1997) Neutral and nonneutral mutations: the creative mix—evolution of complexity in gene interaction systems [published erratum appears in J Mol Evol (1997) 44(4):470]. J Mol Evol 44:S2–S8