Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Drift and Conservatism in RNA Virus Evolution: Are They Adapting or Merely Changing?

Monica Sala and Simon Wain-Hobson

INTRODUCTION

There is no such thing as a perfect machine. Accordingly, nucleic acid polymerization is inevitably error-prone. Yet the notoriety and abundance of RNA viruses attests to their great success as intracellular parasites. Indeed some estimates suggest that 80% of viruses have RNA genomes. It follows that replication without proofreading can be a successful strategy. There is a price to pay, however. Manfred Eigen was the first to point out that without proofreading there is a limit on the size of RNA genomes. Obviously, if the mutation rate is too high, any RNA virus will collapse under mutation pressure. As it happens, RNA viral genomes are up to 32 kb long while mutation rates are 1–2 per genome per cycle or less.

Possibly, RNA viruses and retroviruses have simply not invested in proofreading, in which case mutations represent an inevitable genetic noise, to be tolerated or eliminated. Hence there would be no loss of fitness, fixed mutations being neutral. A corollary of this would be that the intrinsic life style of a virus is set in its genes. The alternative is to suppose that most fixed mutations are beneficial to the virus in allowing it to keep ahead of the host and/or host population. By this token variation is an integral part of the viral modus vivendi. The twin requirements of a successful virus are replication and transmission. Under the rubric replication, a virus could vary to increase its fitness, exploit different target cells or evade adaptive immune responses. In terms of transmission, variation might allow a virus to overcome herd immunity.

These two scenarios emphasize the two sides of the molecular evolution debate; one highlights neutrality while the other puts a premium on positive selection. Purifying or negative selection is ever operative – a poor replicon invariably goes asunder. Through rounds of error and trial, positive selection is the only means of creating a novel replicon. So long as the ecological niche occupied doesn’t change, the virus doesn’t need to change, purifying selection being sufficient to ensure existence. This raises an important issue: we know that, over the time that we are living and loving, as well as doing experiments, writing papers and reviewing, humans are not evolving. Ernst Mayr noted that “the brain of 100 000 years ago is the same brain that is now able to design computers” (Mayr, 1997). Positive fitness selection among mammals is effectively inoperative over our lifetimes. And certainly since we have known about HIV and AIDS.
How is it that vertebrates, invertebrates, plants, fungi and bacteria, all species with a low genomic mutation rate, can control viruses which mutate so much faster – sometimes by a factor of $10^6$ (Holland et al., 1982; Gojobori and Yokoyama, 1985; Domingo et al., 1996). Yet they do. We come to the basic question – to what extent is genetic variation exploited by an RNA virus, if at all? And if so, what is the virus adapting to? The answer invariably given to the second question is “the adaptive immune system” (Seibert et al., 1995).

Yet apart from the vertebrates none of the other groups mentioned above mounts antigen-specific immune responses. This chapter will argue that most fixed mutations are neutral.

**MOLECULAR CLOCKS**

In molecular evolution one of the remarkable observations has been the uniformity of the molecular clock. Although there has been intense debate as to what molecular clocks mean and quite how far they deviate from null hypotheses, fibronectin fixes mutations faster than alpha- or beta-globin, which do so faster than cytochrome c, etc. Rates of amino acid fixation are intrinsic to different proteins. Yet some viruses give rise to persistent infections, others to sequential acute infections. All succumb to the vagaries of transmission bottlenecks. How many rounds of infection are necessary to fix mutations? For example, the tremendous dynamics of viral replication have been described. Whether it be HIV, HBV or HCV, plasma viral turnover is of the order of $10^6-10^9$ virions per day (Ho et al., 1995; Wei et al., 1995; Nowak et al., 1996; Zeuzem et al., 1998).

Between 10% and 90% of plasma virus is cleared. In the case of HIV this can involve more than 200 rounds of sequential replication per year (Wain-Hobson, 1993a,b; Ho et al., 1995; Pelletier et al., 1995; Wei et al., 1995).

Many of these variables and unknowns can be removed by comparing the fixation of amino-acid substitutions in pairs of viral proteins from two genomes. If one assumes that the two gene fragments remain linked, through the hellfire of

![FIGURE 6.1](image_url)

**FIGURE 6.1** Two different representations of distances between proteins encoded by genomes A, B, C, D, E and F. A. Separate phylogenetic analysis based on homologous sequences for proteins y and x. B. Amino acid divergence for pairs of protein y taken from two genomes plotted against divergence for protein x from the same two genomes. See footnote to Table 6.1.
immune responses and bottlenecking inherent in transmission, relative degrees of fixation should be attainable. Note that, so long as frequent recombination between highly divergent genomes is not in evidence, this assumption should be valid. This procedure is outlined in Figure 6.1.

The first example is taken from the vast primate immunodeficiency virus database (LANL, 1998). When normalized to the p66 reverse transcriptase product designated RT, amino acid sequence divergence for p17 Gag, p24 Gag, integrase, Vif, gp120, the ectodomain of gp41 and Nef all reveal highly significant linear relationships (Figure 6.2, Table 6.1). The relative rates vary by a factor of two or more. Why the hyper-variable gp120 protein shows a relatively low degree of change with respect to the reverse

| TABLE 6.1 Smooth drift in pairwise amino acid sequence divergence |
|---------|-------|------------------|---|
| Viruses | Paired proteins | n   | Linear relationship   |
|---------|-----------------|-----|----------------------|
| HIV-1, -2, SIV | p17/RT | 22  | $y = 1.15x - 0.02$ | 0.98* |
| HIV-1, -2, SIV | p24/RT | 22  | $y = 0.91x + 0.03$ | 0.96* |
| HIV-1, -2, SIV | Integrase/RT | 22  | $y = 0.94x - 0.01$ | 0.99* |
| HIV-1, -2, SIV | gp41/RT | 22  | $y = 1.16x + 0.01$ | 0.96* |
| HIV-1, -2, SIV | gp120/RT | 22  | $y = 1.50x + 0.02$ | 0.96* |
| HIV-1, -2, SIV | vif/RT | 22  | $y = 1.49x + 0.01$ | 0.97* |
| HIV-1, -2, SIV | nef/RT | 22  | $y = 1.19x + 0.12$ | 0.94* |
| Rhinovirus | 3C/3D | 5   | $y = 1.05x - 0.01$ | 0.98* |
| Rhinovirus | P2A/3D | 5   | $y = 1.20x - 0.04$ | 0.91+ |
| Rhinovirus | P2C/3D | 5   | $y = 1.14x - 0.02$ | 0.99* |
| Rhinovirus | VP1/3D | 5   | $y = 1.27x + 0.02$ | 0.99* |
| Rhinovirus | VP2/3D | 5   | $y = 0.77x + 0.03$ | 0.97" |
| Rhinovirus | VP3/3D | 5   | $y = 0.96x + 0.02$ | 0.95" |
| Potyvirus | P1/Cl | 21  | $y = 1.42x + 0.15$ | 0.94* |
| Potyvirus | HC-Pro/Cl | 21  | $y = 1.26x - 0.01$ | 0.99* |
| Potyvirus | P3/Cl | 21  | $y = 1.53x + 0.06$ | 0.97* |
| Potyvirus | N1a/Cl | 21  | $y = 1.13x + 0.00$ | 0.99* |
| Potyvirus | N1b/Cl | 21  | $y = 0.87x + 0.01$ | 0.99* |
| Potyvirus | CP/Cl | 21  | $y = 0.92x - 0.01$ | 0.98* |
| Inoviridae | Protein I/II# | 15  | $y = 1.06x + 0.03$ | 1 |
| Inoviridae | Protein III/II# | 15  | $y = 1.02x + 0.11$ | 0.91* |
| Inoviridae | Protein IV/II# | 15  | $y = 1.02x + 0.05$ | 0.99* |

Orthologous protein sequences were multiply aligned and divergence was calculated by the Clustalw program using protein weight matrices of the Blosum series (Henikoff and Henikoff, 1993). It is well established that protein sequence comparisons are more informative when weighted for genetic and structural biases in amino acid replacements. In the Blosum weight matrices series, the actual matrix that was used depends on how similar the sequences to be aligned are. Different matrices work differently at each evolutionary distance. For a given virus, different protein sequence sets were compared to a given reference such as RT in the case of HIV/SIV. n indicates the number of independent two-by-two comparisons. The data were checked for the possibility that a rogue genome strongly influenced the data. Only in the case of the Inoviridae were there insufficient complete sequences, six in fact, to yield satisfying analyses. Instead all pairwise comparisons were made, hence the data points reflect dependent data (#). The form of the linear regressions are given where y and x refer to the first and second protein listed in the column “Paired proteins”. The correlation coefficients r were highly significant in all cases, the corresponding probabilities being: + < 0.02; † < 0.005; * < 0.001.
FIGURE 6.2 Graphical representation of paired divergence for orthologous proteins taken from complete HIV-1, HIV-2 and SIV genome sequences. y = different proteins, x = p66 sequence of the reverse transcriptase (RT). x and y values correspond to Blosum-corrected fractional divergence. Only non-overlapping regions were taken into account. The straight lines were obtained by linear regression analysis. Their characteristics are given in Table 6.1.
transcriptase (RT) can be explained by gap stripping, which eliminates the hypervariable regions. Consequently the gp120 data effectively reflects the conserved regions. The linearity, even out to considerable differences, indicates that multiple substitutions and back mutations, which must be occurring, do so to comparable degrees. Although these data were derived from completely sequenced primate immunodeficiency viral genomes, analyses on larger data sets, such as p17 Gag/p24 Gag or gp120/gp41, yielded relative values that differed from those given in Table 6.1 by at most 14%. The absence of points far from the linear regression substantiates the assumption that recombination between highly divergent genomes is rare. This does not preclude recombination between closely related genomes.

The linear regressions passed close to the origin in nearly all cases. Only for Nef was there some deviation, suggesting that Nef was saturating to a different extent from all other proteins. However, as linear correlations involving Nef data were always statistically significant, this trend may be fortuitous. Note that the data cover the earliest phase, intrapatient variation (generally <10%), continuing smoothly to cover interclade, intertype and finally interspecies

![Graphical representation of paired divergence for orthologous proteins taken from complete rhinovirus genome sequences.](image)

**FIGURE 6.3** Graphical representation of paired divergence for orthologous proteins taken from complete rhinovirus genome sequences. $y =$ different proteins, $x =$ RNA-dependent RNA polymerase (3D). $x$ and $y$ values correspond to Blosum-corrected fractional divergence. The straight lines were obtained by linear regression analysis. Their characteristics are given in Table 6.1.
comparisons. 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. The same forces were uppermost during all stages of diversification. It is remarkable that the very different proteins, such as gp120 and the gp41 ectodomain (surface glycoproteins), p17 Gag and p24 Gag (structural), RT and integrase (enzymes) and Nef and Vif (cytoplasmic), all yield linear relationships (Figure 6.2), as though fixation was an intrinsic property of the protein.

Applying the same analysis to complete rhinoviral genomes yielded comparable results, i.e. highly significant linear relationships for VP1, VP2 and VP3 (capsid proteins), P2A and 3C (proteases), P2C (cytoplasmic proteins involved in membrane reorganization) compared with the RNA-dependent RNA polymerase (3D) as reference (Figure 6.3, Table 6.1). Hence Figure 6.2 does not represent some quirk of primate lentiviruses. Of course, vertebrate viruses have a redoubtable adversary in the host adaptive immune system. The swiftness of secondary responses is reminder enough.

An analysis of proteins derived from complete potyvirus genomes, positive-stranded RNA viruses, yielded highly significant linear relationships (Table 6.1). A number of revealing points can be made. Firstly, the linear relationships hold out to very large Blosum distances (0.9). Secondly, potyviruses infect a wide variety of different plants, as their florid names betray. Finally, the linear relationships cannot result from adaptive immune pressure because plants are devoid of adaptive immune systems. They only have powerful innate immune responses. Nevertheless, there are insufficient insect RNA viral sequences to allow a comparable study. However, a glance at a few beetle nodavirus capsid sequences (Dasgupta et al., 1984; Dasgupta and Sgro, 1989) shows extensive genetic variation with a majority of synonymous base substitutions, typical of most comparisons of mammalian viral sequences (see below). For the time being there doesn’t seem to be anything obviously different about insect virus sequence variation. Although insects do not mount adaptive immune responses, the breadth and complexity of their innate immune systems is salutary (Brey and Hultmark, 1998).

A final example is afforded by the inoviruses, bacteriophages of the fd group, which includes M13. Although DNA viruses fix mutations at a slower rate than RNA viruses, they too show linear relationships among comparisons of their I, II, III and IV proteins (Table 6.1). And of course bacteria are devoid of adaptive immunity as well.

Whether the comparisons were between capsid proteins versus enzymes, or secretory versus cytoplasmic molecules, significant linear relationships were obtained for pairwise comparisons in amino acid variation in all cases. Such proteins are vastly different in their three-dimensional folds and functions. Some are “seen” by humoral immunity, others are not. For the plant viruses and bacteriophages, only innate immunity is operative. It is as though the rate of amino acid sequence accumulation is an intrinsic feature of the protein, reminiscent of the differing slopes for the accumulation of substitutions by alpha-globin and cytochrome c already alluded to. Of course pairwise comparisons of these two proteins from differing organisms would yield a straight line going through the origin in a manner typical of Figures 6.2 and 6.3. Hence it is fairly safe to assume that, for viral proteins too, amino acid substitutions are accumulated smoothly over time. Indeed, this has been shown explicitly for a number of proteins from a varied group of viruses, including the influenza A, coronaviruses, HIV and herpes viruses (Hayashida et al., 1985; Gojobori et al., 1990; Querat et al., 1990; Villaverde et al., 1991; Elena et al., 1992; Sanchez et al., 1992; McGeoch et al., 1995; Yang et al., 1995; Leitner et al., 1997; Plikat et al., 1997).

The above analyses indicate that viral protein diversification is essentially a smooth process, the major parameter being the nature of the protein more than the ecological niche it finds itself in. The simplest hypothesis to explain the smoothness of protein sequence diversification is that the majority of fixed amino acid substitutions are neutral, being accumulated at rates intrinsic to each protein. This is not to say that
positive selection is inoperative, merely that the majority of fixed substitutions are essentially neutral, so much so that it does not strongly distort the data from a linear relationship expected for genetic drift. In other words, neither the impact of different environments nor the ferocity of the adaptive immune response has much to do with fixation of most substitutions. This is important for the one-dimensional man in all of us sequencers who see all mutations and ask questions about genotype and phenotype – usually about genotype.

A short aside is necessary here. It is interesting that in a few areas of RNA virology much has been made of escape from the adaptive immune response, particularly cytotoxic T lymphocytes, so leading to persistence (Nowak and McMichael, 1995; McMichael and Phillips, 1997). However, it is not at all obvious that this is the case (Wain-Hobson, 1996). It must not be forgotten that it is possible to vaccinate against a number of RNA viruses such as measles, polio and yellow fever. Be that as it may, many DNA viruses, intracellular bacteria and parasites persist. In these cases de novo genetic variation arising from point mutations is too slow a means to thwart an adaptive immune response. For example, after 1700 generations, under experimental conditions whereby Muller's ratchet was operative, S. typhimurium accumulated mutations such that only 1% of the 444 lineages tested had suffered an obvious loss of fitness (Anderson and Hughes, 1996). That this number of generations could be achieved within as little as 45 days gives an idea of the time necessary to generate a mutation affecting fitness. This is more than enough time to make a vigorous immune response. Some inklings of immune system escape for the herpes virus EBV (de Campos-Lima et al., 1993, 1994) came to nought (Burrows et al., 1996; Khanna et al., 1997). When antigenic variation is in evidence among DNA-based microbes, it invariably results from the use of cassettes and multicopy genes rather than point mutations resulting from DNA replication. And of course such complex systems could have only come about by natural selection.

Finally de novo genetic variation of an RNA virus has never been suggested or shown to be necessary for the course of an acute infection. For a virus to persist thanks to genetic variation the phenomenon of epitope escape must be strongly in evidence by the time of seroconversion, generally 5–6 weeks. Yet such data are not forthcoming, and not for want of trying. When viruses do play tricks with the immune system it is invariably by way of specific viral gene products that interfere with the mechanics of adaptive and innate immunity (Ploegh, 1998). In the clear cases where genetic variation is exploited by RNA viruses, it is used to overcome barriers to transmission set up by the host population, e.g. herd immunity. The obvious example is influenza A virus antigenic variation in mammals.

Ka/Ks RATIOS

Another way of assessing the contribution of positive selection to sequence variation is to compare the relative proportions of synonymous (Ks) and non-synonymous (Ka) base substitutions per site. A Ka/Ks ratio of less than 1 indicates that purifying selection is uppermost, while a ratio more than 1 is taken as evidence of an excess of positive selection. Comparisons for HIV proteins from different isolates have yielded the same result (Myers and Korber, 1994). Some mileage was made out of the fact that this ratio increased with increasing distance of SIVs with respect to HIV-1, which in turn led to a discussion of SIV pathogenesis (Shpaer and Mullins, 1993). However, this may reflect a lack of adequate correction for multiple hits. This effect is illustrated by a comparison of the set of 72 orthologous proteins encoded by herpes simplex viruses 1 and 2 (HSV-1 and HSV-2; Figure 6.4A). The more divergent the protein sequence, the greater the Ks/K ratio. That some proteins fix substitutions faster than others is no surprise. Yet as Figure 6.4B shows, the Ks values change little as they are near to saturation. When Ks is small, Ks > K. This suggests that reliable interpretation of Ka/Ks ratios is possible only when the degree of nucleic acid divergence is small.

Now this is the realm of viral quasispecies rendered accessible by PCR. HIV studies abound, reflecting both the phenomenal degree
of sequence variation and its importance as a pathogen, so we’ll stick to some such examples that are illustrative. Concerning \(K_a/K_s\) ratios for HIV gene segments, widely varying conclusions have been published supporting all sides (Meyerhans et al., 1989, 1991; Pelletier et al., 1995; Wolinsky et al., 1996; Leigh Brown, 1997; Price et al., 1998), so much so that three comments are in order.

Firstly, many studies have used small numbers of sequences and substitutions and even regions as small as nonameric HLA class-I-restricted epitopes. In such cases statistical analyses are essential to test the significance of the distribution of synonymous and non-synonymous substitutions. This is particularly important as the point substitution matrix is highly biased (Pelletier et al., 1995; Plikat et al., 1997). It turns out that when the proportions are so analysed the distributions are rarely significantly different from the neutral hypothesis (Leigh Brown, 1997). Secondly, the method for counting substitutions is highly variable, ranging from two-by-two comparisons, scoring the number of altered sites in a data set, to phylogenetic reconstruction. This latter method reflects more closely the process of genetic diversification. When so analysed, almost all of the data sets indicated proportions of synonymous to non-synonymous substitutions indistinguishable from that suggested by genetic drift and/or purifying selection (Pelletier et al., 1995; Plikat et al., 1997).

Thirdly, prudence is called for. The fact that obviously defective sequences can be identified, occasionally accounting for large fractions of the sample (Martins et al., 1991; Gao et al., 1992), indicates that not all genomes have undergone the rigours of selection (Nietfield et al., 1995). Indeed, in peripheral blood, HIV is invariably lurking as a silent provirus within a resting memory T-cell. Such T-cells have half lives of 3 months or more (Michie et al., 1992). Hence it would be erroneous to interpret findings based on a single or clustered samples (Price et al., 1997). Only when the above caveats are borne in mind is there any hope of discerning how HIV accumulates mutations. When these issues are attended to, purifying selection is dominant (Pelletier et al., 1995; Leigh Brown, 1997; Plikat et al., 1997).

One must not deny that positive selection is operative, merely that it is hard to pinpoint when looking at full-length sequences. Indeed it is like looking for the proverbial needle in a haystack. In the context of \(K_a/K_s\)-type analyses, the two classic cases in the literature are the HLA class I and II molecules and influenza A haemagglutinin (Hughes, 1998; Hughes and Nei, 1988; Ina and Gojobori, 1994). The peptide contact residues of both class I and II molecules have been under tremendous positive selection. Changes in the five antigenic sites on the flu A haemagglutinin help the virus overcome herd immunity set up during previous flu epidemics.

**FIGURE 6.4** Comparison of \(K_s\) and \(K_a\) values for 72 orthologous proteins encoded by herpes simplex viruses types 1 and 2 (HSV-1 and HSV-2; Dolan et al., 1998). A. \(K_a/K_s\) ratio as a function of uncorrected percentage amino acid sequence divergence (linear regression was \(K_a/K_s = 1.25\) divergence + 0.04, \(r = 0.87\) (\(p < 0.001\))). B. Individual \(K_a\) and \(K_s\) variation with percentage divergence (\(K_a = 0.53\) divergence + 0.35 and \(K_s = 0.76\) divergence - 0.03 with correlation coefficients of 0.54 and 0.97 respectively, \(p < 0.001\) for both). Note how at small degrees of divergence, \(K_s\)>>\(K_a\) decreases as divergence increases. Basically, \(K_s\) is approaching saturation, being uncorrected for multiple and/or back mutations.
Consequently, finding $K_a/K_s > 1$ in these regions was, in some ways, a pyrrhic victory because the papers needed experimental data to identify the positively selected segments in the first place.

More recently Endo et al. (1996) have screened the sequence data bases for proteins in which $K_a/K_s > 1$. Of 3595 homologous gene groups screened, covering about 20,000 sequences, only 17 groups came up positive, of which two were encoded by RNA viruses – the equine infectious anaemia virus envelope proteins and the reovirus σ1 (outer capsid) proteins. The former case is intriguing as there is no obvious correlation between sequence changes and neutralizing antibodies (Carpenter et al., 1987). The authors noted that, when a comparable $K_a/K_s$ analysis was restricted to small segments, the number of protein groups scoring positive rose to 5% (Endo et al., 1996). Despite the explanatory power of these ratios, the number of identifiable cases of positively selected segments is small indeed. These numbers would probably shrink were phylogenetic reconstruction used.

To summarize the section, synonymous changes are invariably more frequent than non-synonymous changes. Positive selection may be operative in the evolution of viral protein sequences. When it is, it apparently exploits only a small fraction of mutants.

**DIVERSITY CALCULATIONS – VIROLOGICAL MAYHEM**

The two rates touted by evolutionary-minded virologists are the mutation rate and the mutation fixation rate. The first describes the rate of genesis of mutations, the second attempts to describe their fixation within the population sampled over a period of time. In the case where all substitutions are neutral, the mutation rate ($m$) equals the fixation rate ($f$) per round of replication. It appears that such a situation applies to the evolution of parts of the SIV and HIV-1 genomes over 1–3 years (Pelletier et al., 1995; Plikat et al., 1997). If fixation rates are measured over one year, then $f = n\cdot m$, where $n$ is the annual number of consecutive rounds of replication. It is simple to show that several hundred rounds of sequential replication are required (Wain-Hobson, 1993b; Pelletier et al., 1995). Given that the proviral load of an HIV-1-positive patient ($=10^{10}$) changes by less than a factor of 10 over 5 years or more, and given the assumption that an infected cell produces sufficient virus to generate two productively infected cells, then annual production would be something akin to $2^{300}$, or $10^{90}$, which is impossible. Clearly even a productive burst size of 2 is too large (Wain-Hobson, 1993a,b). This must be reduced to 1.1 to achieve a realistic proviral load ($1.1^{20} = 10^9$).

Note that the real value for the effective burst size must be even lower, as proviral load is turning over more slowly than once a day. Yet to explain the temporal increase in proviral load, the productive burst size must be 2 or more. Thus the calculation reveals massive destruction of infected cells, precisely what was to be expected from immensely powerful innate and adaptive immune responses.

When purifying selection is in evidence, some additional factor must be introduced to couple the fixation and mutation rates. As the accumulation of most substitutions proceeds in a protein-specific linear manner for small degrees of divergence, the above equation can be modified to $f = P\cdot n\cdot m$, where $1 > P > 0$ is a constant indicating the degree of negative selection. Note immediately that, as $P < 1$, more rounds of replication are needed to produce the same percentage amino acid fixation. A corollary is an even greater degree of destruction of infected cells. Consider the example of a virus that is fixing substitutions only slowly, about $10^{-5}$ per site per year, something like the Ebola virus glycoprotein. The mutation rate for Ebola is not known but is probably around $10^{-4}$ per site per cycle (Drake, 1993). Hence $P\cdot n = 10^{-4}$. What is the value of $n$? Most mammalian viruses replicate within 24 h, while obviously outside of a body they do not replicate. Consequently a value of $n = 50–200$ is probably not unreasonable. Accordingly $P\cdot n = 10^{-3}$ to $5 \times 10^{-4}$. This means that most mutations generated are deleterious. Of those that are fixed, most are neutral, as has been discussed above.
**VIRAL CONSERVATISM**

The last two sentences describe a profoundly conservative strategy – RNA viruses are seen merely to replicate far more than giving rise to genetically distinct, even exotic, siblings. What a stultifying picture, in contrast to the shock-horror of tabloid newspaper virology and that atmospheric, yet profoundly ambiguous term, emerging viruses.

Conservative perhaps, but is there any suggestion that viruses are more or less so than other replicons? Like extrapolation, choosing examples can be problematic. However let’s consider one example, the eukaryotic and retroviral aspartic proteases (Doolittle et al., 1989). The former exist as a monomer with two homologous domains, while the retroviral counterpart functions as a homodimer. Despite these differences the folding patterns are almost identical, meaning that the enzymes may be considered orthologous. Between humans and chickens there is approximately 38% amino acid divergence among typical aspartic proteases (Figure 6.5). The HIV-1 and HIV-2 proteases differ by a little more, 52%. No one would doubt the considerable differences in design, metabolism and lifestyle separating us and chickens. On either side of the HIV protease coding region one finds differences: HIV-1 is \( vpx^vpu \) while HIV-2 is the opposite, i.e. \( vpx^vpu \); there are differences in the size and activities of the \( tat \) gene product; the LTRs are subtly different. Yet both replicate in the same cells in vivo, produce the same disease, albeit with different kinetics: HIV-2 infection progresses more slowly. If these differences are esteemed too substantial, consider the 28% divergence between the HIV and chimpanzee SIV proteases. These two viruses are isogenic. Pig and human chromosomal aspartic proteases may differ by around 17%, the differences between these two species being, George Orwell apart, obvious to all. Even by this crude example, the AIDS viruses would seem to be more conservative than mammals in their evolution. The same argument pertains to the rhinoviral P2A and 3C serine proteases (Figure 6.5).

This conclusion is even more surprising when it is realized that HIV is fixing mutations at a rate of \( 10^{-3} \)–\( 10^{-2} \) per base per year. By contrast, mammals are fixing mutations approximately one million times less rapidly, i.e. approximately \( 10^{-8} \)–\( 10^{-9} \) per base per year (Gojobori and Yokoyama, 1987). However, the generation times of the two are vastly different, about 1 day for HIV and about 15–25 years for humans. Normalizing for this yields a 100-fold higher fixation rate per generation for HIV than for humans. Amalgamating this with the preceding paragraph, we see that HIV is not only evolving qualitatively in a conservative manner, but it is doing so despite a 100-fold greater propensity to accommodate change. The same arguments go...
for almost all RNA viruses and retroviruses. Why is this? Although they mutate rapidly, their hosts are effectively invariant in an evolutionary sense. Probably sticking to the niche is all that matters, which is no mean task given the strength of innate and adaptive antiviral immune responses.

**MANY WAYS TO BE A VIRUS**

John Maynard Smith's argument was simply put. For organisms with a base substitution rate of less than 1 per genome per cycle, he reasoned that all intermediates linking any two sequences must be viable, otherwise the lineage would go extinct. The example used was self explanatory: WORD → WORE → GORE → GONE → GENE (Maynard Smith, 1970). The same is true for viruses, even though their mutation rates are 6 orders higher; the rate for a given protein is still less than 1 substitution per cycle. Even for rather stable viruses like Ebola/Marburg and human T-cell leukaemia virus type 1 and 2 (HTLV-1/-2), the number of intermediates is huge. While the enormity of sequence space is basically impossible to comprehend, the amount accessible to a virus remains vast. For the lineage to exist, the probability of finding a viable mutant must be at least 1/population size within the host.

Imagine a stem-loop structure. Any replacement of a G:C base pair must proceed by a single substitution, given that the probability of a double mutation is approximately $10^{-4}$ that of a single mutation. Let substitution of a G:C pair pass by a G:U intermediate, finishing up as A:U. Although G:U mismatches are the most stable of all mismatches, they are less so than either a G:C or an A:U pair. There are two scenarios: either the G:U substitution is of so little consequence that it is fixed *per se*, in which case there would be no selection pressure to complete the process to A:U. Alternatively, the G:U substitution is sufficiently deleterious for selection of a secondary mutation to occur from a pool of variants, so completing the process. Yet the G:U intermediate cannot be so debilitating otherwise the process would have little chance of going to completion. Note also that if the fitness difference is small with respect to the G:C or A:U forms, more rounds of replication are necessary to achieve fixation of G:U to A:U. A corollary is that there must be a range within which fitness variation is tolerated. This is reminiscent of nearly neutral theories of evolution and their extension to RNA viruses (Chao, 1997; Ohta, 1997). Note also that from a theoretical perspective the same secondary structure can be found in all parts of sequence space with easy connectivity (Schuster, 1995; Schuster *et al.*, 1997).

Figure 6.6 shows a number of variations on an HIV stem-loop structure, crucial for ribosomal frameshifting between the *gag* and *pol* open reading frames. There have been substitutions at positions 1, 2, 5, 8 and 11 and even an opening up of the loop. All come from viable strains, yet the environment in which these structures are operative, the human ribosome, is invariant. If the changes are all neutral the situation is formally comparable to the steady accumulation of amino acid substitutions. However, if the intermediates are less fit, it has to be understood how they can survive long enough in the face of a plethora of competitors, approximately 1/mutation rate or about 10,000 for HIV. The latter is probably the case as there are HIV-1 genomes with C:G to U:A substitutions at positions 5 and 8 (Figure 6.6). Extensions of nearly neutral theory would fit these findings well (Chao, 1997; Ohta, 1997). That there are many solutions to this stem-loop problem is clear. If HIV-2 is brought into the picture, the remarkable plurality of solutions is further emphasized (Figure 6.6).

Degeneracy in solutions found by viruses is revealed by some interesting experiments on viral revertants. The initial lesions substantially inactivated the virus. Yet with a bit of patience, sometimes more than 6 months, replication-competent variants that were not back mutations were identified (Klaver and Berkhout, 1994; Olsthoorn *et al.*, 1994; Berkhout *et al.*, 1997; Willey *et al.*, 1988; Escarmis *et al.*, 1999). As the frequencies of mutation and back mutation are not equivalent, such findings are, perhaps, not surprising. What they show is the range of possible solutions adjacent to that created by the experimentalist. Loss of fitness can be achieved
FIGURE 6.6 “Shifty” RNA stem-loop structures from HIV-1 M, N and O group strains as well as from HIV-2 Rod. This structure is part of the information that instructs the ribosome to shift from the gag open reading frame to that of pol. In addition to the hairpin is a heptameric sequence (underlined). Frameshifting occur within the gag UUA codon within the heptamer and continues AGG.GAA etc.* highlights differences in nucleotide sequences compared with the M reference strain LAI.

by sequential plaquing of RNA viruses, the so-called Müller’s ratchet experiment, which has been analysed at the genetic level for FMDV (Escarmis et al., 1996). Different lesions characterized different lineages. Recent work was aimed at characterizing the molecular basis of fitness recovery following large population passage. Not one solution was found but a variety, even in parallel experiments (Escarmis et al., 1999). This reveals the impact of chance in fitness selection on a finite population of variants, which is trivially small given the immensity of sequence space.

Another example of degeneracy in viable solutions is the isolation of functional ribozymes from randomly synthesized RNA (Bartel and Szostak, 1993; Ekland et al., 1995). From a pool of approximately $10^{14}$ variants, through repeated rounds of positive selection, it was estimated that the frequency of the ribozyme was of the order of $10^{-8}$, which is small indeed. Yet $10^{-8} \times 10^{14} = 10^6$. Even erring by four orders of magnitude, 100 distinct ribozymes could well have been present in the initial pool. Although the sequence space occupied may well represent a tiny proportion of that possible for a RNA molecule of length $n$, the space is so large that the number of viable solutions is large, large enough to permit a plethora of parallel solutions to the same problem. These experiments, ribozyme from dust, are cases in plurality.
Further evidence of the large proportion of viable solutions in protein sequence space comes from in vitro mutagenesis. For example bacteriophage T4 lysozyme can absorb large numbers of substitutions (Rennell et al., 1991) with very few sites resisting replacement (Figure 6.7). Other examples include the lymphokine, interleukin 3, in which some forms with enhanced characteristics were noted (Olins et al., 1995; Klein et al., 1997). With modern mutagenic methods allowing mutation rates of 0.1 per base per site or less, hypermutants of the E. coli R67 dihydrofolate reductase (DHFR) were found by random sequencing of as little as 30 clones (Martinez et al., 1996). Whatever the mutation bias, mutants with 3–5 amino acid replacements within the 78-residue protein could be attained (Figure 6.8). Other mutagenesis studies sought enzymes with enhanced catalytic constants or chemical stability. For subtilisin E variants with enhanced features for two parameters could be identified from a relatively small population of randomly mutagenized molecules (Kucher and Arnold, 1997). These data indicate that functional sequence space is probably far more dense than hitherto thought.

Most of the above examples concern maintenance or enhancement of function. An interesting example was recently afforded by engineering cyclophilin into a proline-specific endopeptidase (Quéméneur et al., 1998). The proline binding pocket of cyclophilin was modified such that a single amino acid change (A91S) generated a novel serine endopeptidase with a $10^{10}$ mol/l proficiency with respect to cyclophilin. Addition of two further substitutions (F104H and N106D) generated a serine–aspartic-acid–histidine catalytic triad, the hallmark of serine proteases. The final enzyme proficiency was $3.5 \times 10^9$ mol/l, typical of many natural enzymes. This shows the interconnectedness of sequence spaces for two functionally very different proteins. If sequence space were sparsely populated, the probability of observing such phenomena would be small.

Many viruses recombine, and via molecular biology more can be made, some of which are tremendously useful research tools, such as the SHIVs, chimeras between SIV and HIV (Figure 6.9). Although many groups have tried to recombine naturally HIV-1 with HIV-2 or SIV, none has succeeded. Natural and artificial recombination represent major jumps in sequence space. That one can observe such genomes means that the new site in functional sequence space must be only a few mutations

![FIGURE 6.7](image-url) Systematic amino acid replacement of bacteriophage T4 lysozyme residues. Amber stop codons were engineered singly into each residue apart from the initiator methionine. The plasmids were used to transform 13 suppressor strains. Of the resulting 2015 single amino acid substitutions, 328 were found to be sufficiently deleterious to inhibit plaque formation. More than half (55%) of the positions in the protein tolerated all substitutions examined. The side chains of residues that were refractory to substitution were generally inaccessible to solvent. The catalytic residues are Glu11 and Asp20. Adapted from Rennell et al., 1991.
from a reasonably viable solution, otherwise it would take too long to generate large numbers of cycles and, along with them, mutants. The ferocity of innate and adaptive immunity must never be forgotten.

Off on an apparent tangent, the phylogeny of Geoffrey Chaucer’s *The Canterbury Tales* was recently analysed by programs tried and tested for nucleic acid sequences. The authors used 850 lines from 58 fifteenth-century manuscripts (Barbrook et al., 1998). Apart from the fact that it appears that Chaucer did not leave a final version but some annotated working copy, the radiation in medieval English space is fascinating. All the versions are viable and “phenotypically” equivalent even though the “genotypes” are not so. It is ironic that William Caxton’s first printed edition was far removed from the original. (N.B., printers merely make fewer errors than scribes, tantamount to adding a 3’ exonuclease domain to an RNA polymerase).

Given the inevitability of mutation, is it possible that over the aeons natural selection has selected for proteins that are robust, those that are capable of absorbing endless substitutions? For if amino acid substitutions were very difficult to fix, huge populations would need to be explored before change could be accommodated. Recently the unstructured N-terminal segment of the *E. coli* R67 DHFR was shown to stabilize amino acid substitutions in a non-functional miniprotein devoid of this segment (Figure 6.10; Martinez et al., 1996). While the mechanism by which this occurs is unknown, it suggests that there may be parts of proteins, even multiple or discontinuous segments, that may help the protein accommodate inevitable change. Formally it can be seen that such

---

**FIGURE 6.8** A modest collection of hypermutated dihydrofolate reductase protein sequences encoded by the *E. coli* R67 plasmid. All were trimethoprim resistant. Only differences with respect to the parent sequence are shown. A representation of the three-dimensional structure is shown above. Adapted from Martinez et al., 1996, with permission.
FIGURE 6.9 Genetic organization of naturally occurring HIV-1 and SIV recombinants and unnatural, genetically engineered, SIV-HIV-1 chimeras called SHIVs. Segments are hatched according to strain origin. References are HIV-1 Mal and HIV-1 IBNG (Gao et al., 1996), HIV-1 92RW009.6 (Gao et al., 1998), SIVagm sab-1 (Jin et al., 1994) and SHIVsbg (Dunn et al., 1996).
proteins would have both short- and long-term selective advantages, for they would permit the generation of larger populations of relatively viable variants as well as buffering the lineage against the effects of bottlenecks.

**A HANDFUL OF RESIDUES**

What fraction of amino acid residues is necessary for function? Answer – very few. A few examples taken from among the primate immunodeficiency viruses are typical. Almost all these viruses infect the same target cell using the membrane proteins CD4 and CCR5. Primary HIV-1 isolates use the chemokine receptor CCR5 and rarely the homologous molecule CXCR4, which differs by 81% in its extracellular domains. Yet two substitutions in the viral envelope protein gp120 are sufficient to allow use of the CXCR4 molecule (Hwang et al., 1991). Curiously, the CCR5 chemokine receptor homologue, US28, encoded by human cytomegalovirus, can be used by HIV-1 despite the fact that US28 and CCR5 differ by 88% in the same extracellular regions (Pleskoff et al., 1997).

Clearly only a small set of residues are necessary for docking (Brelot et al., 1997).

Another example is afforded by the Vpu protein, which is unique to HIV-1 and the chimpanzee virus SIVcpz (Huet et al., 1990). Vpu is a small protein inserted into the endoplasmic reticulum, tucked well away from humoral immunity. Despite an average amino acid sequence difference of 0.5% among orthologous human and chimpanzee proteins, HIV/SIVcpz Vpu divergence is almost beyond reliable sequence alignment (Figure 6.11): an N-terminal hydrophobic membrane anchor and a couple of perfectly conserved serine residues, which are phosphorylated, and that’s about it. Among HIV-1 strains, or between SIVcpz sequences, the situation was a little better. Yet the necessity of keeping Vpu is beyond doubt. A fine final example concerns the HIV/SIV Rev proteins. These small nuclear proteins are crucial to viral replication. Despite this, only 5 residues are perfectly conserved. The situation has been taken beyond the limit, at least ex vivo, in that the HTLV-1 Rex protein can functionally complement for HIV-1 Rev (Rimsky et al., 1989), despite the fact that they are completely different proteins.

The above is reminiscent of what is known about enzymes and surface recognition. Provided the protein fold is maintained, only a small fraction of residues actually contribute to function, a point made recently in two reviews on RNA viral proteases (Ryan and Flint, 1997; Ryan et al., 1998). Insertions and deletions are generally less than 2–3 residues in length and
FIGURE 6.11 A selection of Vpu protein sequences from HIV-1 M, N and O group viruses as well as the two SIV chimpanzee (cpz) strains. Hyphens were used to maximize sequence alignments. Numbers 1 and 2 indicate the number of different amino acids at that position for the HIV-1 separately and HIV-1 and SIV cpz sequences together.

confined to turns, loops and coils (Pascarella and Argos, 1992). If globular proteins or at least domains are, to a first approximation, taken as spheres, then the surface area is the least for any volume. If amino acids are equally viewed as smaller, closely packed spheres, then a minimum number will be exposed on the surface, ready to partake in recognition and function.

ARE VIRUSES OPTIMIZED?

The molecular biologist frequently thinks like an engineer who can redesign from scratch. Yet replicons have been constrained by a series of historical events representing variations on a founding theme. While they are fit enough to survive, are they the best possible? This question is salutary, for we live in a society that is more and more competitive and, thanks to global communications, knows about the most successful athletes or businessmen worldwide. Yet who can remember the name of any Olympic athlete who came in fourth? Is not fourth best in any large population remarkable?
finding suggest that the optimum is actually away from mass production? Yet if there is a shadow of reality in this thesis, it indicates that fitness optima in vivo may not necessarily parallel the expectations of fitness based on ex-vivo models. In this context note also that HTLV-1 infects exactly the same cell as HIV, yet its codon usage is very different from that of HIV and the Thy-1 gene (Seiki et al., 1983).

If fitness optimization were ever operative in vivo, then one would predict steady increases in virulence for those viruses that do not set up herd immunity. At some point a plateau would be reached. Yet the higgledy-piggledy way by which virulent strains come and go suggests that this is not so. Some might use the word stochastic. Whatever. If fitness selection can be overridden and we don’t have a good theory for it, then we’re in a sorry state.

THE (DIFFICULT) TRANSITION FROM EX VIVO TO IN VIVO

There is abundant evidence that, as a good first approximation, RNA viruses ex vivo perform as expected from the quasi-species model (Holland et al., 1982; Eigen and Biebricher, 1988; Duarte et al., 1992; Clarke et al., 1993; Eigen, 1993; Novella et al., 1995; Domingo et al., 1996; Quer et al., 1996; Domingo and Holland, 1997), which is fitness dominated. Problems arise transposing it to the in vivo situation, notably:

- First and foremost: how does one determine fitness in vivo? Should such measurements score intrahost viral titres or transmission probabilities from an index case? (If a virus doesn’t spread it’s dead.) For outbred populations, is it in fact virulence?
- Second: host innate immunity is hugely powerful, a fact leading Rolf Zinkernagel to remark with typical aplomb that in terms of immunity “an interferon receptor knock-out mouse is a 1% mouse” (Huang et al., 1993; van den Broek et al., 1995a,b). Yet the enhanced susceptibility of SCID humans or various knock-out mice to infections indicates the part played by adaptive immunity.

For example, influenza A can persist in SCID children (Rocha et al., 1991). How are innate and adaptive immune responses coupled and how are they influenced by genetic polymorphisms?

- Third: with acquired immunity rising by day 3 in an acute infection, the virus is replicating in the face of a predator whose amplitude is increasing.
- Fourth: immune responses are density-dependent. That is, the more the virus replicates the stronger the immune response. If the relationship were simply linear one could see how a virus might be able to keep just ahead, given a short lag in the immune response time. But if it were non-linear? Indeed it must be so, otherwise it would not be possible to resolve an acute infection. It is not easy to discern where optimal viral fitness would lie.
- Fifth: The wrath of combined immune responses is such that there is massive viral turnover. For the three best known cases, HIV, HBV and HCV, between 10^8 and 10^12 virions are turning over daily, representing between 10% and 90% of the whole (Ho et al., 1995; Wei et al., 1995; Nowak et al., 1996; Zeuzem et al., 1998). Indeed, these are probably underestimates, given beautiful data from the late 1950s and 1960s showing that, for a variety of RNA viruses, plasma titres decay with a half-life of 15–20 min, whether the animal be immunologically naive or primed (Mimms, 1959; Nathanson and Harrington, 1967). From this one may conclude that any viral population is unlikely to be in equilibrium. And if a population is not in equilibrium, fitness selection is compromised.
- Sixth: A glance at any histology slide or textbook is a salient reminder of spatial discontinuities over distances of one or two cell diameters. For example the hugely delocalized immune system is characterized by a multitude of different lymphoid organs, a myriad of subtly different susceptible cell types, and a mêlée of membrane molecules. The exquisite spatial heterogeneity of HIV within the epidermis and splenic white pulps has been described (Cheynier et al., 1994,
The same seems to be true for HCV-infected liver (Martell et al., 1992). For HPV infiltration of skin, spatial discontinuities and gradients are also apparent. Discontinuities reduce the possibilities for competition and hence selection of the fitter forms. Indeed the Müller’s ratchet experiment and clonal heterogeneity are the most vivid expressions of this.

- Seventh: much has been made of privileged sites and viral reservoirs. Basically this is reminding us of the fact that immune surveillance is modulated in some organs like the brain. There are some suggestions that cytotoxic T-cells have difficulty infiltrating the kidney. Viral reservoirs undermine fitness selection.

- Eighth: in the case of the immunodeficiency viruses, antigenic stimulation of infected yet resting memory T-cells means that variants may become amplified for reasons that have nothing to do with the fitness of the variant (Cheynier et al., 1994, 1998).

Given the non-equilibrium structure of viral variants, vastly restricted population sizes in respect to sequence space, founder effects in vivo take on great importance. While answers for some of these issues seem far away, constraints on fitness selection cannot be so strong that a chain of infections becomes a Müller’s ratchet experiment. Yet is that correct? In the experiments with phage Φ6, VSV and FMDV, most of the lineages resulted in decreased fitness. Yet for some there were no changes, while for a few there were even increases in the fitness vectors (Chao, 1990; Duarte et al., 1992; Escarmis et al., 1996). Could symptomatic infections reflect bottleneck transmission of those fitter clones with asymptomatic (subclinical) infection representing fitness-compromised clones?

Analysis of RNA viruses ex vivo is analogous to the study of bacteria in chemostats. Fitness selection dominates. Yet there is a world of difference between bacterial strains so selected and natural isolates. One of the observations frequently made upon isolation of pathogenic bacteria is the loss of bacterial virulence determinants (Miller et al., 1989). Indeed, ex-vivo passage of RNA viruses has been used to select for attenuated strains used in vaccination.

Mayr again: “Wherever one looks in nature, one finds uniqueness” (Mayr, 1997). As mentioned, the cardinal difference between the behaviour of RNA viruses ex vivo and in vivo is the existence of spatial discontinuities. For replicons, cloning is the ultimate separation. It allows a variant to break away from dominating competitors, disrupts or uncouples a fitter variant locked in competitive exclusion (de la Torre and Holland, 1990). The effect of bottlenecking on fitness, as well as the Müller’s ratchet experiments, have been described (Chao, 1990; Duarte et al., 1992; Novella et al., 1995; Escarmis et al., 1996). Transmission frequently involves massive bottlenecking, and is very much an exercise in cloning. All this should not surprise because allopatric speciation is omnipresent in the origin of species, Darwin’s Galapagos finches being an obvious example.

AN ORIGIN OF VIRAL SPECIES?

A virus must replicate sufficiently within a host to permit infection of another susceptible host. If the new host is of the same species, differences between the two are minimal – a small degree of polymorphism being inevitable in outbred populations. Given that viruses with a small coding capacity interact particularly intimately with the host-cell machinery, it follows that infection of a host from a related species has a greater probability of succeeding if the cellular machinery is comparable. Indeed, the closer the two species, the greater the probability. In turn, if the virus gets a toehold and can generate a quasispecies, then only few mutations would probably be necessary to adapt to the new niche.

Yet species is a difficult word. What might a viral species be? Martin (1993) wrote a fascinating review on the number of extinct primate species estimated from the fossil record. Depending on the emergence time of primates of modern aspect, he was able to estimate the total number that existed as 5500–6500. The present number of 200 primates species would thus
represent about 3.4–3.8%. More importantly from our viewpoint was his calculation of the average survival time of fossil primate species as a mere 1 million years (Martin, 1993). Given that RNA viruses are fixing mutations approximately 1 million times faster than mammals (Holland et al., 1982; Gojobori and Yokoyama, 1985; Domingo et al., 1996), a viral species would become extinct after approximately 1 year! Immediately the annual influenza A strain comes to mind. Yet rabies, polio and HTLV-1 have arguably been around for millennia. Clearly the word "species", when taken from primatology, cannot apply to the viral world. Frogs provide a more interesting example. They have been around for several hundred millions of years, and members of some lineages can interbreed despite 75 million years separation. Naturally, their protein sequences have not stood still during that time (Wilson et al., 1977). Enough is conserved to allow breeding. Maybe the primate picture has undue weight in our appreciation of virology. Phenotype can be maintained despite changes in genotype – obvious to a biologist.

As usual, Holland wasn't far from the mark when he wrote:

As human populations continue to grow exponentially, the number of ecological niches for human RNA virus evolution grows apace and new human virus outbreaks will likely increase apace. Most new human viruses will be unremarkable – that is they will generally resemble old ones. Inevitably, some will be quite remarkable, and quite undesirable. When discussing RNA virus evolution, to call an outbreak (such as AIDS) remarkable is merely to state that it is of lower probability than an unremarkable outbreak. (Holland et al., 1992)

New viruses can and do emerge but on a scale that is probably 15–20 logs less than the number of viral mutants generated up to that defining moment (Wain-Hobson, 1993). They will result from a small number of mutations and a dose of reproductive isolation.

CONCLUSION

The above has attempted to show that the vast majority of genetic changes fixed by RNA viruses are essentially neutral or nearly neutral in character. Positive selection exploits a small proportion of genetic variants, while functional sequence space is sufficiently dense, allowing viable solutions to be found. Although evolution has connotations of change, what has always counted is natural selection or adaptation. It is the only force for the genesis of a novel replicon. Once adapted to its niche, there is no need to change. In such circumstances an RNA

| 100 bp | 1 kb | 10 kb | 100 kb | 1 Mb | 10 Mb |
|-------|------|------|-------|------|------|
| viroids | RNA viruses |
| retroviruses | DNA viruses |
| bacteria, archaea, yeast |

FIGURE 6.12 Latitude in microbial genome sizes. RNA viruses and retroviruses are confined to one log variation in size (3 to = 32 kb). By contrast, DNA viruses span more than 2.5 logs going from the single-stranded porcine circovirus (1.8 kb) to chlorella virus (~330 kb, encoding at least 12 DNA endonuclease/methyltransferase genes; Zhang et al., 1998) and bacteriophage G (~670 kb). The distinction between phage DNA and a plasmid has often proven difficult (Waldor and Mekalanos, 1996). As can be seen, the genome size of the largest DNA viruses overlaps the smallest intracellular bacteria such as mycoplasmas (580 and 816 kb) and is not too far short of autonomous bacteria such as Haemophilus influenzae (1.83 Mb).
virus would no longer be adapting, even though it could be changing.

Why is the evolution of RNA viruses so conservative? Why do they mutate rapidly yet remain phenotypically stable? The lack of proofreading proscribes the genesis of large genomes, restricting their genome sizes to a 1 log range (Figure 6.12). Among the smallest RNA and retroviruses are MS2 and hepatitis B virus, both about 3 kb, while the largest are the coronaviruses at 32 kb or more. Most of their proteins are structural or regulatory and take up the largest part of the coding capacity of the virus. Additional proteins broadening the range of interactions with the host cell, or rendering the replicon more autonomous, are relatively few. Large, gene-sized duplications that may contribute to diversification and novel phenotypes are rare, reducing the exploration of new horizons. Thus, evolution of RNA viruses is probably conservative because they cannot shuffle domains so generating new combinations.

That the information capacity of RNA viral genomes is limited by a lack of proofreading is neither here nor there, for they are remarkably successful parasites. RNA viruses change far more than they adapt.

ACKNOWLEDGEMENTS

We would like to thank past and present members of the laboratory and numerous colleagues for endless discussions over the years. Mark Mascolini needs a special word of thanks for painstakingly going through the manuscript. This laboratory is supported by grants from the Institut Pasteur and the Agence Nationale pour la Recherche sur le SIDA.

REFERENCES

Andersson, D.I. and Hughes, D. (1996) Muller's ratchet decreases fitness of a DNA-based microbe. Proc. Natl Acad. Sci. USA, 93, 906–907.

Andre, S., Seed, B., Eberle, J., Schraut, W., Bultmann, A. and Haas, J. (1998) Increased immune response elicited by DNA vaccination with a synthetic gp120 sequence with optimized codon usage. J. Virol., 72, 1497–1503.

Barbrook, A.C., Howe, C.J., Blake, N. and Robinson, P. (1998) The phylogeny of The Canterbury Tales. Nature, 394, 839.

Bartel, D.P. and Szostak, J.W. (1993) Isolation of new ribozymes from a large pool of random sequences. Science, 261, 1411–1418.

Berkhout, B., Klaver, B. and Das, A.T. (1997) Forced evolution of a regulatory RNA helix in the HIV-1 genome. Nucleic Acids Res., 25, 940–947.

Brelot, A., Heverker, N., Pleskoff, O., Sol, N. and Alizon, M. (1997) Role of the first and third extracellular domains of CXCR-4 in human immunodeficiency virus coreceptor activity. J. Virol., 71, 4744–4751.

Brey, P.T. and Hultmark, D. (1998) Molecular Mechanisms of Immune Responses in Insects. Chapman & Hall, London.

Bronson, E.C. and Anderson, J.N. (1994) Nucleotide composition as a driving force in the evolution of retroviruses. J. Mol. Evol., 38, 506–532.

Burrows, J.M., Burrows, S.R., Poulsen, L.M., Sculley, T.B., Moss, D.J. and Khanna, R. (1996) Unusually high frequency of Epstein–Barr virus genetic variants in Papua New Guinea that can escape cytotoxic T-cell recognition: implications for virus evolution. J. Virol., 70, 2490–2496.

Carpenter, S., Evans, L.H., Sevoian, M. and Chesbro, B. (1987) Role of host immune response in selection of equine infectious anemia virus variants. J. Virol., 61, 3783–3789.

Chao, L. (1990) Fitness of RNA virus decreased by Muller’s ratchet. Nature, 348, 454–455.

Chao, L. (1997) Evolution of sex and the molecular clock in RNA viruses. Gene, 205, 301–308.

Cheynier, R., Henrichwark, S., Hadida, F. et al. (1994) HIV and T-cell expansion in splenic white pulps is accompanied by infiltration of HIV-specific cytotoxic T-lymphocytes. Cell, 78, 373–387.

Cheynier, R., Gratton, S., Halloran, M., Stahmer, I., Letvin, N.L. and Wain-Hobson, S. (1998)
Antigenic stimulation by BCG as an in vivo driving force for SIV replication and dissemination. Nat. Med., 4, 421–427.

Clarke, D.K., Duarte, E.A., Moya, A., Elena, S.F., Domingo, E. and Holland, J. (1993) Genetic bottlenecks and population passages cause profound fitness differences in RNA viruses. J. Virol., 67, 222–228.

Dasgupta, R. and Sgro, J.Y. (1989) Nucleotide sequences of three Nodavirus RNA2’s: the messengers for their coat protein precursors. Nucleic Acids Res., 17, 7525–7526.

Dasgupta, R., Ghosh, A., Dasmahapatra, B., Guarino, L.A. and Kaesberg, P. (1984) Primary and secondary structure of black beetle virus RNA2, the genomic messenger for BBV coat protein precursor. Nucleic Acids Res., 12, 7215–7223.

Doman, A., Jamieson, F.E., Cunningham, C., Barnett, B.C. and McGeoch, D.J. (1998) The genome sequence of herpes simplex virus type 2. J. Virol., 72, 2010–2021.

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

Domingo, E., Escarmis, C., Sevilla, N. et al. (1996) Basic concepts in RNA virus evolution. FASEB J., 10, 859–864.

Doolittle, R.F., Feng, D.F., Johnson, M.S. and McClure, M.A. (1989) Origins and evolutionary relationships of retroviruses. Q. Rev. Biol., 64, 1–30.

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

Duarte, E., Clarke, D., Moya, A., Domingo, E. and Holland, J. (1992) Rapid fitness losses in mammalian RNA virus clones due to Muller’s ratchet. Proc. Natl Acad. Sci. USA, 89, 6015–6019.

Dunn, C.S., Beyer, C., Kiency, M.P. et al. (1996) High viral load and CD4 lymphopenia in rhesus and cynomolgus macaques infected by a chimeric primate lentivirus constructed using the env, rev, tat, and vpu genes from HIV-1 Lai. Virology, 223, 351–361.

Eigen, M. (1993) The viral quasispecies. Sci. Am., 269, 42–49.

Eigen, M. and Biebricher, C.K. (1988) Sequence space and quasispecies distribution. In: RNA Genetics, vol. 3, (eds Domingo, E., Holland, J.J. and Ahlquist, P.), pp. 211–245. CRC Press, Boca Raton, FL.

Ekland, E.H., Szostak, J.W. and Bartel, D.P. (1995) Structurally complex and highly active RNA ligases derived from random RNA sequences. Science, 269, 364–370.

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

Endo, T., Ikeo, K. and Gojobori, T. (1996) Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol., 13, 685–690.

Escarmis, C., Davila, M., Charpentier, N., Bracho, A., Moya, A. and Domingo, E. (1996) Genetic lesions associated with Muller’s ratchet in an RNA virus. J. Mol. Biol., 264, 255–267.

Escarmis, C., Davila, M. and Domingo, E. (1999) Multiple molecular pathways for fitness recovery of an RNA virus dibilitated by operation of Muller’s ratchet. J. Mol. Biol., 285, 495–505.

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

Gao, F., Yue, L., White, A.T. et al. (1992) Human infection by genetically diverse SIV-sm related HIV-2 in West Africa. Nature, 358, 495–499.

Gao, F., Robertson, D.L., Morrison, S.G. et al. (1996) The heterosexual human immunodeficiency virus type 1 epidemic in Thailand is...
caused by an intersubtype (A/E) recombinant of African origin. J. Virol., 70, 7013–7029.

Gao, F., Robertson, D.L., Carruthers, C.D. et al. (1998) A comprehensive panel of near-full-length clones and reference sequences for non-subtype B isolates of human immunodeficiency virus type 1. J. Virol., 72, 5680–5698.

Gojobori, T. and 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. and Yokoyama, S. (1987) Molecular evolutionary rates of oncogenes. J. Mol. Evol., 26, 148–156.

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

Haas, J., Park, E.C. and Seed, B. (1996) Codon usage limitation in the expression of HIV-1 envelope glycoprotein. Curr. Biol., 6, 315–324.

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

Henikoff, S. and Henikoff, J.G. (1993) Performance evaluation of amino acid substitution matrices. Proteins, 17, 49–61.

Ho, D.D., Neumann, A.U., Perelson, A.S., Chen, W., Leonard, J.M. and Markowitz, M. (1995) Rapid turnover of plasma virions and CD4 lymphocytes in HIV-1 infection. Nature, 373, 123–126.

Holland, J., Spindler, K., Horodyski, F., Grabau, E., Nichol, S. and Vande Pol, X. (1982) Rapid evolution of RNA genomes. Science, 215, 1577–1585.

Holland, J.J., de le Torre, J.C. and Steinhauer, D.A. (1992) RNA virus populations as quasi-species. Curr. Topics Microbiol. Immunol., 176, 1–20.

Huang, S., Hendriks, W., Althage, A. et al. (1993) Immune response in mice that lack the interferon-gamma receptor. Science, 259, 1742–1745.

Huet, T., Cheynier, R., Meyerhans, A., Roelants, G. and Wain-Hobson, S. (1990) Genetic organization of a chimpanzee lentivirus related to HIV-1. Nature, 345, 356–359.

Hughes, A.L. (1998) Protein phylogenies provide evidence of a radical discontinuity between arthropod and vertebrate immune systems. Immunogenet., 47, 283–296.

Hughes, A.L. and Nei, M. (1988) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature, 335, 167–170.

Hwang, S.S., Boyle, T.J., Lyerly, H.K. and Cullen, B.R. (1991) Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. Science, 253, 71–74.

Ina, Y. and Gojobori, T. (1994) Statistical analysis of nucleotide sequences of the hemagglutinin gene of human influenza A viruses. Proc. Natl Acad. Sci. USA, 91, 8388–8392.

Jin, M.J., Hui, H., Robertson, D.L. et al. (1994) Mosaic genome structure of simian immunodeficiency virus from West African green monkeys. EMBO J., 13, 2935–2947.

Khanna, R., Burrows, S.R. and Burrows, J.M. (1997) The role of cytotoxic T-lymphocytes in the evolution of genetically stable viruses. Trends Microbiol., 5, 64–69.

Klaver, B. and Berkhout, B. (1994) Evolution of a disrupted TAR RNA hairpin structure in the HIV-1 virus. EMBO J., 13, 2650–2659.

Klein, B.K., Feng, Y., McWherter, C.A., Hood, W.F., Paik, K. and McKearn, I.P. (1997) The receptor binding site of human interleukin-3 defined by mutagenesis and molecular modeling. J. Biol. Chem., 272, 22630–22641.

Kucher, O. and Arnold, F.H. (1997) Directed evolution of enzyme catalysts. Trends Biotechnol., 15, 523–530.

LANL (1999) http://hiv-web.lanl.gov.

Leigh Brown, A.J. (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. and 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.

McGeoch, D.J., Cook, S., Dolan, A., Jamieson, F.E. and Telford, E.A. (1995) Molecular phylogeny and evolutionary timescale for the family of mammalian herpesviruses. J. Mol. Biol., 247, 443–458.
McMichael, A.J. and Phillips, R.E. (1997) Escape of human immunodeficiency virus from immune control. Ann. Rev. Immunol., 15, 271–296.

Martell, M., Esteban, J.I., Quer, J. et al. (1992) Hepatitis C virus (HCV) circulates as a population of different but closely related genomes: quasispecies nature of HCV genome distribution. J. Virol., 66, 3225–3229.

Martin, R.D. (1993) Primate origins: plugging the gaps. Nature, 363, 223–234.

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

Martins, L.P., Chenciner, N., Asjo, B., Meyerhans, A. and Wain-Hobson, S. (1991) Independent fluctuation of human immunodeficiency virus type 1 rev and gp41 quasi-species in vivo. J. Virol., 65, 4502–4507.

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

Mayr, E. (1997) This is Biology. Belknap Press, Cambridge, MA.

Meyerhans, A., Cheynier, R., Albert, J. et al. (1989) Temporal fluctuations in HIV quasi-species in vivo are not reflected by sequential HIV isolations. Cell, 58, 901–910.

Meyerhans, A., Dadaglio, G., Vartanian, J.P. et al. (1991) In vivo persistence of a HIV-1-encoded HLA-B27-restricted cytotoxic T-lymphocyte epitope despite specific in vitro reactivity. Eur. J. Immunol., 21, 2637–2640.

Meyerhans, A., Vartanian, J.P., Hultgren, C. et al. (1994) Restriction and enhancement of human immunodeficiency virus type 1 replication by modulation of intracellular deoxynucleoside triphosphate pools. J. Virol., 68, 535–540.

Michie, C.A., McLean, A., Alcock, C. and Beverley, P.C. (1992) Lifespan of human lymphocyte subsets defined by CD45 isoforms. Nature, 360, 264–265.

Miller, J.F., Mekalanos, J.J. and Falkow, S. (1989) Coordinate regulation and sensory transduction in the control of bacterial virulence. Science, 243, 916–922.

Mimms, C.A. (1959) The response of mice to large intravenous injections of ectromelia virus. I. The fate of injected virus. J. Exp. Path., 40, 533–542.

Myers, G. and Korber, B. (1994) The future of human immunodeficiency virus. In: The Evolutionary Biology of Viruses, (ed. Morse, S.S.), pp. 211–232. Raven Press, New York.

Nathanson, N. and Harrington, B. (1967) Experimental infection of monkeys with Langat virus. Am. J. Epidemiol., 85, 494–502.

Nietfeld, W., Bauer, M., Fevrier, M. et al. (1995) Sequence constraints and recognition by CTL of an HLA-B27-restricted HIV-1 gag epitope. J. Immunol., 154, 2189–2197.

Novella, I.S., Elena, S.F., Moya, A., Domingo, E. and Holland, J.J. (1995) Size of genetic bottlenecks leading to virus fitness loss is determined by mean initial population fitness. J. Virol., 69, 2869–2872.

Nowak, M.A. and McMichael, A.J. (1995) How HIV defeats the immune system. Sci. Am., 273, 58–65.

Nowak, M.A., Bonhoeffer, S., Hill, A.M., Boehme, R., Thomas, H.C. and McDade, H. (1996) Viral dynamics in hepatitis B virus infection. Proc. Natl Acad. Sci. USA, 93, 4398–4402.

Ohta, T. (1997) The meaning of near-neutrality at coding and non-coding regions. Gene, 205, 261–267.

Olins, P.O., Bauer, S.C., Braford-Goldberg, S. et al. (1995) Saturation mutagenesis of human interleukin-3. J. Biol. Chem., 270, 23754–23760.

Olsthoorn, R.C., Licis, N. and van Duin, J. (1994) Leeway and constraints in the forced evolution of a regulatory RNA helix. EMBO J., 13, 2660–2668.

Pascarella, S. and Argos, P. (1992) Analysis of insertions/deletions in protein structures. J. Mol. Biol., 20, 461–471.

Pelletier, E., Saurin, W., Cheynier, R., Letvin, N.L. and 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.

Pleskoff, O., Treboute, C., Brelot, A., Heveker, N., Seman, M. and Alizon, M. (1997) Identification of a chemokine receptor encoded by human cytomegalovirus as a cofactor for HIV-1 entry. Science, 276, 1874–1878.

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

Ploegh, H.L. (1998) Viral strategies of immune evasion. Science, 280, 248–253.

Price, D.A., Goulder, P.J., Klennerman, P. et al. (1997) Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. Proc. Natl Acad. Sci. USA, 94, 1890–1895.

Price, D.A., Sewell, A.K., Dong, T. et al. (1998) Antigen-specific release of beta-chemokines by anti-HIV-1 cytotoxic T lymphocytes. Curr. Biol., 8, 355–358.

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

Quenat, G., Audoly, G., Sonigo, P. and 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, S.E., Hardy, L.W. and Poteete, A.R. (1991) Systematic mutation of bacteriophage T4 lysozyme. J. Mol. Biol., 222, 67–88.

Rimsky, L., Dodon, M.D., Dixon, E.P. and Greene, W.C. (1989) Trans-dominant inactivation of HTLV-I and HIV-1 gene expression by mutation of the HTLV-I Rex transactivator. Nature, 341, 453–456.

Rocha, E., Cox, N., Black, R.A., Harmon, M.W., Harisson, C.J. and Kendall, A. (1991) Antigenic and genetic variation in influenza A (H1N1) virus isolates recovered from a persistently infected immunodeficient child. J. Virol., 65, 2340–2350.

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

Ryan, M.D., Monaghan, S. and Flint, M. (1998) Virus-encoded proteinases of the Flaviviridae. J. Gen. Virol., 79, 947–959.
and-mouth disease virus. Can quasispecies define a transient molecular clock? *Gene*, **103**, 147–153.

Wain-Hobson, S. (1993a) The fastest genome evolution ever described: HIV variation *in situ*. *Curr. Opin. Genet. Devel.*, **3**, 878–883.

Wain-Hobson, S. (1993b) Viral burden in AIDS. *Nature*, **366**, 22.

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

Wain-Hobson, S., Sonigo, P., Danos, O., Cole, S. and Alizon, M. (1985) Nucleotide sequence of the AIDS virus, LAV. *Cell*, **40**, 9–17.

Waldor, M.K. and Mekalanos, J.J. (1996) Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science*, **272**, 1910–1914.

Wei, X., Ghosh, S.K., Taylor, M.E. et al. (1995) Viral dynamics in human immunodeficiency virus type 1 infection. *Nature*, **373**, 117–122.

Willey, R.L., Smith, D.H., Lasky, L.A. et al. (1988) *In vitro* mutagenesis identifies a region within the envelope gene of the human immunodeficiency virus that is critical for infectivity. *J. Virol.*, **62**, 139–147.

Wilson, A.C., Carlson, S.S. and White, T.J. (1977) Biochemical evolution. *Annu. Rev. Biochem.*, **46**, 573–639.

Wolinsky, S.M., Korber, B.T., Neumann, A.U. et al. (1996) Adaptive evolution of human immunodeficiency virus-type 1 during the natural course of infection. *Science*, **272**, 537–542.

Yang, Z., Lauder, I.J. and Lin, H.J. (1995) Molecular evolution of the hepatitis B virus genome. *J. Mol. Evol.*, **41**, 587–596.

Zeuzem, S., Lee, J.H., Franke, A. et al. (1998) Quantification of the initial decline of serum hepatitis C virus RNA and response to interferon alfa. *Hepatology*, **27**, 1149–1156.

Zhang, Y., Nelson, M., Nietfeldt, J. et al. (1998) Chlorella virus NY-2A encodes at least 12 DNA endonuclease/methyltransferase genes. *Virology*, **240**, 366–375.