J Mol Evol (1984) 20:135–146

Journal of
Molecular Evolution

Divided Genomes and Intrinsic Noise

J. Pressing and D.C. Reanney

Department of Microbiology, La Trobe University, Bundoora, Victoria 3083, Australia

Summary. Segmental genomes (i.e., genomes in which the genetic information is dispersed between two or more discrete molecules) are abundant in RNA viruses, but virtually absent in DNA viruses. It has been suggested that the division of information in RNA viruses expands the pool of variation available to natural selection by providing for the reassortment of modular RNAs from different genetic sources. This explanation is based on the apparent inability of related RNA molecules to undergo the kinds of physical recombination that generate variation among related DNA molecules. In this paper we propose a radically different hypothesis. Self-replicating RNA genomes have an error rate of about 10^{-3}–10^{-4} substitutions per base per generation, whereas for DNA genomes the corresponding figure is 10^{-9}–10^{-11}. Thus the level of noise in the RNA copier process is five to eight orders of magnitude higher than that in the DNA process. Since a small module of information has a higher chance of passing undamaged through a noisy channel than does a large one, the division of RNA viral information among separate small units increases its overall chances of survival. The selective advantage of genome segmentation is most easily modelled for modular RNAs wrapped up in separate viral coats. If modular RNAs are brought together in a common viral coat, segmentation is advantageous only when interactions among the modular RNAs are selective enough to provide some degree of discrimination against miscopied sequences. This requirement is most clearly met by the reoviruses.

Key words: RNA viruses — Divided genomes — Copying fidelity — Intrinsic selection pressure

Offprint requests to: D.C. Reanney

Introduction

Divided genomes, in which the genetic information is dispersed among two or more physically separate molecules, occur in over 17 groups of RNA viruses (Matthews 1979). By contrast, only one group of DNA viruses appears to have a divided genome structure (Haber et al. 1981). Perhaps the most popular explanation for the existence of divided genomes is that they allow modular RNAs from related but different clones to exchange sequences by reassortment (Jaspers 1974; Joklik 1974; Nahmias and Reanney 1977; Reijnders 1978; Lane 1979). According to this view divided genomes have been selected for during evolution because they expand the pool of variation in interbreeding populations. The expanded variation model is widely favoured because reassortment has been documented convincingly in many groups of viruses with divided genomes, e.g., the influenza group (Webster and Granoff 1974; Palese and Young 1982), the reoviruses (Ahmed and Fields 1981; Joklik 1981) and the rotaviruses (Greenberg et al. 1982). Mixed infection experiments also support the concept that related viruses can exchange RNA segments: The RNA-3 of cowpea chlorotic mottle virus can substitute for the RNA-3 of brome mosaic virus (Bancroft 1972), and the Q strain of cucumber mosaic virus (CMV) and the V strain of tomato aspermy virus (TAV) form a vigorous hybrid that contains RNA-3 of CMV and RNA-1 and RNA-2 of TAV (Habili and Francki 1974).

Recent developments, however, have cast very substantial doubt on the concept that the raison d'être of divided genomes is to generate diversity for evolution. For one thing this explanation is strongly 'group selectionist', and the group-selection argument is now invoked only as a last resort by

biologists (see Maynard-Smith 1978; Rose and Doolittle 1983). Perhaps more to the point, the rate of mutation in RNA viruses such as the Qβ is so high that each viable viral genome in a clonally derived population differs from the 'average' sequence of the parental population in one or two positions (Domingo et al. 1978). This appears to be true of all RNA viruses (Holland et al. 1982; Reanney 1984). Thus the identity of any RNA virus genome in nature is only maintained because selection continually removes the unacceptable variants that are continually generated by the error-prone copier mechanism. In a situation in which the pool of preexisting genetic variation is so high that 'only 14% of the population consists of "wild-type" phage' (i.e., virus) (Domingo et al. 1978) the need for additional variation to be generated by reassortment is not obvious. Other explanations may therefore be sought.

Paradoxically, the very thing that damages the credibility of the 'generator of diversity' model, namely the high level of noise in the RNA copier mechanism, provides, in our view, the correct explanation for the widespread occurrence of divided viral genomes.

Noise Levels in RNA and DNA Copier Systems

DNA is usually a double-stranded molecule that replicates in a semiconservative fashion. By contrast, RNA molecules replicate asymmetrically from single strands even if the copying process uses a double-helical template (as in reoviruses). This distinction has the fundamental consequence that lesions in RNA molecules cannot be repaired. This is because known error-correcting mechanisms always use the information specified by one intact strand of a duplex molecule to guide restorative processes on the damaged complementary strand (see Loeb and Kunkel 1982). Because RNA lacks the editing and proofreading functions of DNA, the frequency of mutation in RNA copier systems is between 100,000 and 100,000,000 times greater than that in DNA copier systems (see Kornberg 1980; Holland et al. 1982).

The rate of mutation in the ribophage Qβ has been accurately calibrated at 3×10^{-4} substitutions per nucleotide per generation (Domingo et al. 1978). Studies in polio, Sendai and influenza viruses on the evolution of variants resistant to monoclonal antibodies suggest that this value is essentially the same in all RNA viruses (Portner et al. 1980; Prabhakar et al. 1982). However, measurements of mutation frequency are of dubious value unless the temperature at which replication occurs is taken into account, since error rates in RNA replication increase with temperature (Reanney and Pressing 1983). In the absence of repair, errors will accumulate in the genome. An RNA virus that replicates at 37°C may thus transmit considerably more errors to its progeny than one that replicates at 15°C.

Nondividing RNA molecules also can accumulate errors from a variety of sources. Physical agents may damage RNA by deaminating cytosine to uracil (heat) or inducing pyrimidine-to-pyrimidine dimers (ultraviolet radiation), while a chemical agent such as hydrogen peroxide may generate a variety of changes due to its oxidative capacity. Since these premutational lesions cannot be repaired in an RNA system, they may cause a significant, long-term deterioration in the quality of the genetic information encoded in RNA molecules. RNA genomes are also vulnerable to cleavage by RNases, which are abundant in most sites of RNA virus multiplication.

Collectively these observations suggest that the level of noise in the RNA information transmission mechanism may be much higher than is generally appreciated. How have RNA genomes compensated for these hazards? We suggest that genome segmentation is a direct adaptive consequence of the high error burden placed on RNA genes. This suggestion is based on the observation that a small module of information has a higher chance of passing through a noisy channel without damage than does a large one. Essentially our model depends on the fact that all of the agents that induce errors in RNA (including the replicase mechanism itself) do so in a length-dependent manner.

To provide a rigorous, quantitative model of the 'protective' effect(s) of segmentation we have compared the 'survival rating' of a divided genome with that of an undivided genome of equivalent length. A detailed treatment requires consideration not only of the error rate per generation due to copying, but also of the long-term differential survival rate of mutants with respect to the wild type as a result of processes of chemical equilibrium and kinetics. The details of such processes are still poorly known, and we avoid the need to consider them by building our model in the following way:

Let infection occur via a population of initially error-free viruses. From this base line we then derive the fidelity of the next generation of viruses for both the divided and undivided genome cases. The quotient of these two fidelities is considered to represent the selective advantage, K, of the divided genome strategy in each generation of virus multiplication.

Development of the model shows that the protective effects of genome subdivision differ depending on whether the various modular RNAs are united in one capsid (monocompartment viruses) or dispersed among separate capsids (multicompartment viruses).

Multicompartment Viruses

Consider a simple case where two RNA modules, A and B, are separately replicated and encapsidated. Let q be the mean copying fidelity per nucleotide per generation. The corresponding error rate is then $1 - q$. Copy error evidently makes the largest contribution to $1 - q$, but there is also some deterioration in the quality of the genetic information due to the various factors mentioned earlier. The effects of these latter types of error burden should be comparable and are treated further on in the discussion.

Consider first the undivided genome case in which a fraction r of a total of N virus particles enters the cells. If the genome length is L nucleotides the number of correctly replicated viruses in the subsequent generation is just rNq^L and the overall fidelity of the process is

$$f = rNq^L/N = rq^L \tag{1}$$

For a comparable divided genome, let the lengths of the two modular RNAs A and B be L_A and L_B nucleotides, with $L_A + L_B = L$. Of the initial N_A A-modules, rN_A will enter cells. However, these will not be viable in any given cell unless at least one B is also present. Hence the number of viable As will be $rN_A\lambda_B$, where λ_B is the overall fraction of cells inoculated with RNA B. These viable As will then produce a next generation of $rN_Aq^{L_A}\lambda_B$ correct As. Similarly, the number of correct Bs produced will be $rN_Bq^{L_B}\lambda_A$.

The resulting fidelities of the new generation will differ for A- and B-type modules. To obtain an overall fidelity we reason as follows. The final amount of correct A-type RNA is $rN_Aq^{L_A}\lambda_B \cdot L_A$ and that of B-type RNA $rN_Bq^{L_B}\lambda_A \cdot L_B$, for a total of $r(N_AL_A\, q^{L_A}\lambda_B + N_BL_Bq^{L_B}\lambda_A)$. The initial amount of RNA was $N_AL_A + N_BL_B$. Hence, the overall fidelity is given by

$$F = r \cdot \frac{N_AL_Aq^{L_A}\lambda_B + N_BL_Bq^{L_B}\lambda_A}{N_AL_A + N_BL_B} \tag{2}$$

and K, the selective advantage per generation, by

$$K = \frac{F}{f} = \frac{N_AL_Aq^{L_A}\lambda_B + N_BL_Bq^{L_B}\lambda_A}{q^L(N_AL_A + N_BL_B)} \tag{3}$$

This result may readily be generalized to the case of n particles in which each A is viable only if it enters a cell already inoculated with at least one B, one C, etc. We obtain

$$K = \sum_{i=1}^{n} N_iL_iq^{L_i} \prod_{\substack{j=1 \\ j\neq i}}^{n} \lambda_j \Big/ \left(q^L \sum_{i=1}^{n} N_iL_i\right) \tag{4}$$

where N_i is the initial number of RNA type i, L_i is its genome length, and λ_i is the overall fraction of cells inoculated with RNA i.

Table 1. Size and module number of multicompartment viral RNAs[a]

Virus	No. of modular RNAs	Sizes of modules
Nepo	2	2.4, 1.4–2.2
Pea enation mosaic	2	1.7, 1.3
Como	2	2.0, 1.4
Tobra	2	2.4, 0.6–1.4[b]
Cucumo	3	1.3, 1.1, 0.8
Bromo	3	1.1, 1.0, 0.7
Ilar	3	1.1, 0.9, 0.7
Alfalfa mosaic	3	1.1, 0.8, 0.7
Average		1.27
Mean deviation		0.6–2.4

[a] Data from Matthews (1979). Some virus groups for which information is lacking or imprecise have not been included

[b] The significant size asymmetry of this group is probably related to their highly elongated tubular structure, with consequent capsid size variation (cf. tobacco mosaic virus). This contrasts with the isometric or polyhedral or, in alfalfa mosaic virus bacilliform, capsid symmetries of the other groups

Several comments follow from this equation. First, λ_i is clearly an increasing function of N_i. In fact the most plausible hypothesis, random transmission of particles, may be shown to yield $\lambda_i = 1 - e^{-rN_i/N_H}$, where N_H is the number of host cells. Second, the expression for K may be shown to achieve its maximum value when both numbers and sizes of modules are equally distributed. This is shown mathematically in Appendix 1. The prediction of equal size distribution is well supported by existing data on multicompartment viruses (Table 1), considering that other biochemical factors are bound to influence the distribution to some degree. We expect equal number distribution to hold as well, but data on this are not available.

Taking equal distribution as given, we may assess the dependence of K on n by writing $L_i \approx L/n$, $N_i \approx N/n$ and $\prod_{j\neq i}^{n} \lambda_j \approx \lambda^{n-1}$ to obtain

$$K = \lambda^{n-1}q^{\left(\frac{1}{n}-1\right)L} \tag{5}$$

valid for $n > 1$, where λ is the average fraction of cells infected. From this it may be seen that the selective advantage of genome segmentation increases with error rate $(1-q)$ and genome size (L).

Since $q^{[(1/n)-1]L}$ is a slowly increasing function of n (when $n > 1$) for typical q and L values and λ^{n-1} is sharply decreasing n is unlikely to be large (see Fig. 1). The data are in accord with this prediction: As seen in Table 1, n is never greater than 3. As will be seen below, this contrasts sharply with the monocompartment case.

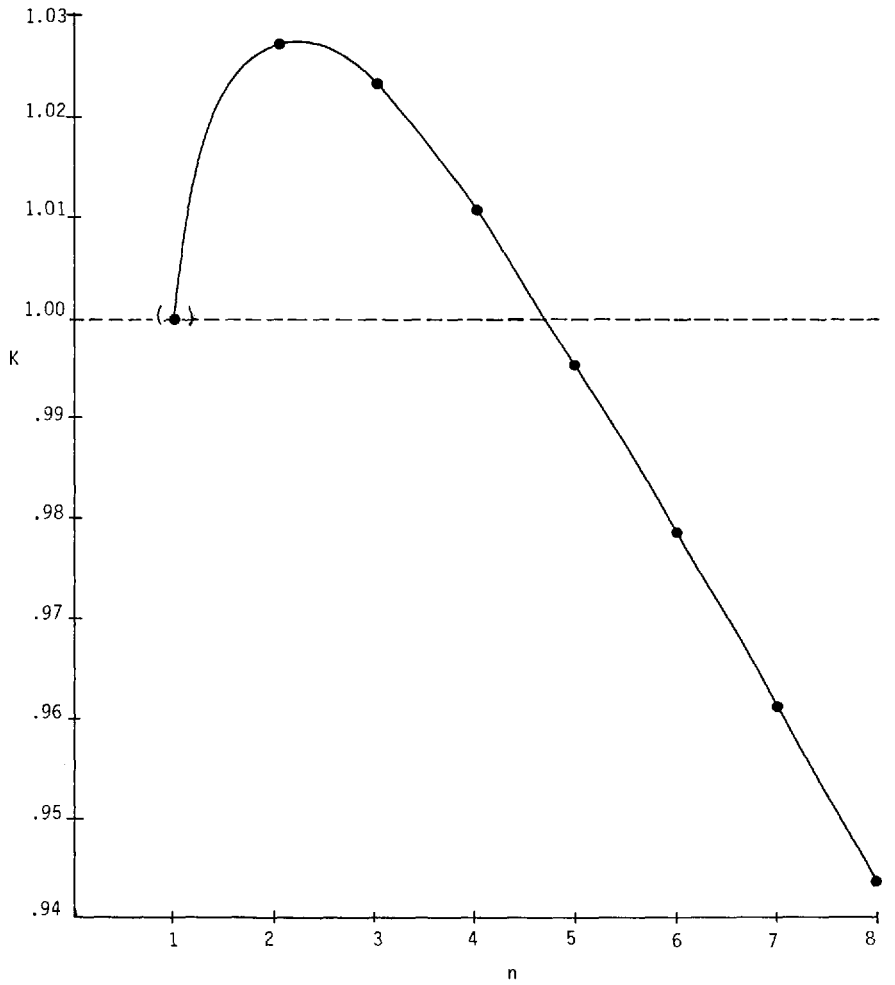

Fig. 1. Variation of relative advantage of divided genome strategy, K, with number of particles, n, for the multicompartment case [see Eq. (5)]. $\lambda = 0.97$; $q^{-L} = 1.10$ (10% errors). The precise range of n within which $K > 1$ is a sensitive function of the infection density λ, and attention should be directed to the shape of the curve rather than its specific n-intercept

Monocompartment Viruses

We consider first the undivided genome case. If a population of initially error-free viruses of genome length L nucleotides reproduces inside hosts, the overall fidelity after one generation will be $f = q^L$ and the fraction of incorrect copies, $1-q^L$.

For a comparable divided genome, consider first the simple case in which there are two modular RNAs A and B of length L_A and L_B nucleotides, with $L_A + L_B = L$. The result of the first replication will be to produce a new generation of N_A As and N_B Bs. We assume here for simplicity that $N_A = N_B$, since (a) this clearly corresponds to the most efficient use of cell material for reproduction and (b) in many cases the mechanism of reproduction is known to be co-operative (e.g., when A codes for the replicase of both A and B and B codes for the protein coat). The fraction of correct copies would then be the same ($q^{L_A} \cdot q^{L_B} = q^L$) as in the undivided genome case if all possible pairings of one A and one B (viz. AB, A*B, AB*, A*B*, where * indicates a miscopied sequence) were equally likely to result in encapsidation.

However, genetic and molecular data indicate that this equiprobability is not achieved in nature. For example, the reoviruses contain 10 to 12 modular RNAs (Matthews 1979). These RNAs are assembled in a highly specific manner such that each virus particle normally accumulates the correct quota of the genetic information (Silverstein et al. 1976; Joklik 1981). The molecular basis of this specificity is believed to be a set of selective RNA:RNA interactions and/or RNA:protein interactions (Silverstein et al. 1976) that operate while the RNAs are single stranded (Joklik 1981). Lane (1979) has proposed a co-operative process in which the binding of one RNA during assembly alters a nucleation complex to create a binding site for a second RNA and so on. If the binding of a given RNA is faulty, the subsequent RNAs have a smaller chance of entering the nascent particle.

These selective interactions constitute a crude form of molecular proofreading (Reanney 1982), since miscopied RNAs are less likely than well-copied RNAs to recognise sequence-specific elements in complementary RNAs or RNA-binding sites in

proteins. This bias towards accurately copied sequences allows us to define a molecular 'discrimination coefficient' σ which theoretically may vary between 0 and 1.0. (When $\sigma = 1$, there is no discrimination against erroneous copies, and when $\sigma = 0$, there is complete discrimination.) It is possible to provide a physical interpretation of σ in the following manner. For the process of RNA association and encapsidation the fundamental rate constant, k, may be assumed to follow an Arrhenius-type equation

$$k = Ce^{-E/kT} \tag{6}$$

where C is a constant, kT is the Boltzmann factor, and E is the activation energy of the reaction A + B → AB. Now consider the reaction A + B* → AB* and let the presence of one error in B* increase the activation energy by ϵ, two errors by 2ϵ and so on. If the mean number of errors in B* is m_B then the corresponding activation energy is $E + m_B\epsilon$ and the corresponding reaction constant is

$$Ce^{-(E+m_B\epsilon)/kT}$$

which can be shown (Appendix 2) to yield

$$\sigma_B = e^{-m_B\epsilon/kT} \tag{7}$$

The result for module A has a similar form.

The precise value of ϵ is not known, but it may be estimated from the difference in mean free energy (ΔG) of hydrogen bonding of an incorrect base pair (e.g., G–A) relative to a correct pair (e.g., G–C) times the probability that this error is located or expressed in the s sites of specific interaction between the A-type and B-type RNAs. A rough estimate is thus

$$\epsilon \approx \frac{s}{L_B}\Delta G \tag{8}$$

Since an average ΔG value is reported to be approximately 1.8 kcal (Tinoco et al. 1971), for sample values of $L_B = 1000$ and $s = 30$ we obtain $\epsilon \approx 0.054$ kcal.[1] For the well-studied Qβ system, in which m is known to be about 1.5, this yields $\sigma = 0.88$ at 37°C. This value agrees well with the measured reaction-rate ratios between mutant and wild-type forms of this virus (Domingo et al. 1978), which were typically 0.8–0.9, and shows that our model is physically realistic.

To see the effect of σ we first note that the A- and B-type RNAs may be equal or different in size. By appropriate choice of labels we then write $L_A \leq L_B$. Since then $q^{L_A} \geq q^{L_B}$, there will either be fewer correct Bs than As or an equal number. In either case the number of correct Bs will limit the number of possible correct encapsidations. Now for the case of random association ($\sigma = 1$) the relative numbers of products would be:

AB	$q^{L_A} \cdot q^{L_B}$
A*B	$(1 - q^{L_A}) \cdot q^{L_B}$
AB*	$q^{L_A} \cdot (1 - q^{L_B})$
A*B*	$(1 - q^{L_B}) \cdot (1 - q^{L_A})$

For $\sigma < 1$ a certain fraction $g(\sigma)$ of the Bs that (for $\sigma = 1$) paired to form A*Bs will now preferentially associate with As that (for $\sigma = 1$) paired to form AB*s. This is the only way additional ABs may be formed. Since the number of A*Bs is less than or equal to the number of AB*s, there will be fewer such Bs than As, or an exactly equal number. Con-

[1] The sample values for L and s have been chosen because they accord with known data. The value of 1000 given for L is a 'rounded off' figure for genome segment 8 of simian rotavirus 11, which has been sequenced and which has a length of 1059 bases (Both et al. 1982). The average length of the 11 modular RNAs of this monocompartment virus is about 1110 bases, according to estimates of genome segment length given in the same reference.

The value for s is difficult to estimate because the mechanism of reovirus assembly is poorly understood. The value must be greater than 10, otherwise it would not be possible to assemble 10 to 12 modular RNAs in the same coat (see Lane 1979). If one assumes the basis of this specificity to be a set of RNA:RNA interactions, then the presumed interaction between small nuclear RNAs and the 'consensus' sequences at the exon:intron junction of split genes may provide a model of what happens. The number of nucleotides in this 'consensus' sequence is about 28 (Rogers and Wall 1980). If, as seems more likely, the specificity resides in RNA:protein interactions one can be more confident, because many examples of specific DNA:protein interactions are known in detail. The average of the published values is about 30, the value for s used by us in this paper. There is no reason to believe that the number of nucleotides recognised by a protein would be greatly different if the substrate were single-stranded RNA, because the high degree of secondary structure in known RNAs confines most bases to double-helical regions.

Size of target sequences for some polynucleotide:protein interactions

Protein	Target-sequence size (no. of nucleotides)	Reference
RNA polymerase	42 (consensus)	Proc Natl Acad Sci USA 80:3203 (1983)
Cap (*E. coli*)	25 (protection by catabolic activator protein from DNase 1)	Proc Natl Acad Sci USA 80:1594 (1983)
Lac repressor (*E. coli*)	25 (protected by repressor)	Science 187:27 (1975)
	26 (deduced)	Proc Natl Acad Sci USA 75:3578 (1978)
RNA polymerase (phage T7)	24	Proc Natl Acad Sci USA 74:4266 (1977)
	Average = 29	

Table 2. Monocompartment viruses: size and segmentation[a]

Virus[b]	No. of modular RNAs	Total mol. wt. ($\times 10^6$ daltons)	Host
Reo[c]	10–12	12–20	Animals, plants
Cysto[c]	3	10.4	Bacteria
Tomato spotted wilt	4	7.5	Plants
Bunya	3	5.5	Animals
Orthomyxo	8	5	Animals
Arena	5 (3 of host origin)	?	Animals
Compare with:			
(a) *Largest continuous RNA genomes ($\times 10^6$ daltons)*			
Corona		5.5–6.1 (8.1?)	
Paramyxo		5–7	
(b) *Average size for multicompartment genomes*			
		2.9	
(c) *Average size for all RNA viral genomes*			
		3–4	

[a] Data from Matthews (1979)
[b] Names recommended by Matthews (1979) have been used, the suffixes "virus" or "viridae" being omitted for simplicity
[c] Double-stranded RNA virus

sequently the number of extra ABs formed for $\sigma < 1$ is proportional to $g(\sigma)\cdot$(number of A*Bs for $\sigma = 1$) $= g(\sigma)(1-q^{L_A})q^{L_B}$, and the total fraction of correct ABs is

$$F = q^{L_A}q^{L_B} + g(\sigma)(1 - q^{L_A})q^{L_B} \tag{9}$$

whence we find

$$K = F/f = 1 + g(\sigma)(q^{-L_A} - 1) \tag{10}$$

where g satisfies the conditions

$$0 \leq g(\sigma) \leq 1, \quad g(0) = 1, \quad g(1) = 0$$
$$\text{and } \frac{dg}{d\sigma} \leq 0 \tag{11}$$

(The first three conditions are necessary boundary conditions and the last implies that fidelity is a monotonic function of σ.)

A general expression for $g(\sigma)$ is not required for the derivation of several results. It is sufficient to note that we expect any variation of $g(\sigma)$ with q to be minimal and to satisfy $dg/d(1-q) > 0$. This is because an increase in error rate $(1-q)$ will increase the mean number of errors per RNA molecule and cause σ to decrease. That is, $d\sigma/d(1-q) < 0$. Therefore,

$$\frac{dg}{d(1-q)} = \frac{dg}{d\sigma}\cdot\frac{d\sigma}{d(1-q)} > 0$$

From equation (10) we then obtain the following important conclusions:

1. The selective advantage (K) of genome segmentation increases with error rate $(1-q)$.

2. As σ approaches 1, K approaches 1 as expected. As σ approaches 0, K approaches q^{-L_A}.

3. The overall advantage is cumulative. Even if the K value per generation is small, the effect is multiplied over succeeding generations. The dependence of K on $g(\sigma)$ may be shown in tabular form, where we assume a high error rate of approximately 10% per module so that $q^{-L_A} = 1.10$, as:

$g(\sigma)$	K
0.05	1.005
0.5	1.05
0.95	1.095

4. The overall effect of segmentation in monocompartment viruses is to decrease by a factor of as much as 2 the amount of genetic information subject to noise-induced damage, since for $\sigma \to 0$ with $L_A \approx L_B$, $K \to q^{-L/2}$.

Maximum message length is inversely related to the frequency of copy error (Eigen and Schuster 1977). Point (4) therefore suggests that genome subdivision should allow a segmented RNA virus to exceed significantly the maximum information capacity of a continuous RNA genome. It is therefore of interest to note that the largest RNA genomes are found among the reoviruses (Table 2), and it can hardly be coincidental that these upper-limit genomes ($12–20 \times 10^6$ daltons) are also the most highly divided (10–12 modules per particle). Indeed, a relatively high degree of segmentation for monocompartment viruses is predicted by our model (see below).

The Dependence of K on n for Monocompartment Viruses

We now turn to a second question: Can we predict an optimal number of modular RNAs for monoparticulate viruses for which the above advantage holds true?

Consider n RNAs A, B, C, etc., and assume as before that we label the largest segment L_B so that the concentration of Bs is a limiting factor. Let a mean σ operate for each RNA association inside the viral particle. The enhancement of correct genomes over incorrect will now entail the factor $g(\sigma)$ (see earlier discussion) for each separate viral association. The fraction of correct copies may then be shown to be:

$$F = (\text{fraction with 0 errors for } \sigma = 1) + g(\sigma)\sum_{i \neq B}^{n}\left(\begin{matrix}\text{fraction with error(s)}\\ \text{in module i}\end{matrix}\right)$$

$$+ g^2(\sigma) \sum_{\substack{i<j \\ i,j\neq B}}^{n\ n} \begin{pmatrix} \text{fraction with error(s)} \\ \text{in modules i and j} \end{pmatrix}$$

$$+ g^3(\sigma) \sum_{\substack{i<j<k \\ i,j,k\neq B}}^{n\ n\ n} \left(\quad \right) + \ldots$$

or

$$F = q^L + g(\sigma) \sum_{i\neq B}^{n} (1 - q^{L_i}) q^{L-L_i}$$

$$+ g^2(\sigma) \sum_{\substack{i<j \\ i,j\neq B}}^{n\ n} (1 - q^{L_i})(1 - q^{L_j}) q^{L-L_i-L_j} + \ldots$$

Using the equation

$$\prod_i^n (1 + x_i) = 1 + \sum_i^n x_i + \sum_{i<j}^{n\ n} x_i x_j + \sum_{i<j<k}^{n\ n\ n} x_i x_j x_k + \ldots$$

and $K = F/q^L$ this gives for the selective advantage in the n-module case:

$$K = \prod_{i\neq B}^{n} [1 + g(\sigma)(q^{-L_i} - 1)] \qquad (12)$$

where the product is calculated over all but the largest modular RNA.

It is not possible to determine unambiguously the optimal size distribution of RNA modules from Eq. (9) for the mathematical reason that the precise variation of $g(\sigma)$ with module size is unknown. If $g(\sigma)$ were independent of module size then K would be maximized for an equal size distribution. It is far more likely that $g(\sigma)$ increases with individual module sizes, since covalent and hydrogen bonding sites sensitive to errors will increase in number as L_i rises [this is readily shown from Eq. (4) of Appendix 2]. In this case the optimal distribution is expected to be asymmetrical; just how asymmetrical cannot be determined without a functional form for $g(\sigma)$. In any case an asymmetrical distribution is strongly preferred in nature for viruses with highly divided genomes, such as the influenza group and the reoviruses (Table 3).

To assess the dependence of K on n it is convenient to simplify the expression using the approximation $q^x = 1 + x(q-1)$ for q values very close to 1. For typical values of q and L involved here, this approximation is quite reasonable. We then find

$$K \approx 1 + g(\sigma)(1-q)(L-L_B)$$

Table 3. Size asymmetry among the modular components of viruses with highly divided genomes

Virus	Module no.	Mol. wt. ($\times 10^6$ daltons)
Fowl plague (orthomyxo) (Bromley and Barry, 1973)	1	1.19
	2	1.02
	3	1.00
	4	0.83
	5	0.68
	6	0.58
	7	0.32
	8	0.28
Influenza (orthomyxo)[a] A(WSN) (HON1) (Palese and Schulman, 1976)	1	1.07
	2	0.95
	3	0.80
	4	0.65
	5	0.60
	6	0.47
	7	0.39
	8	0.34
	9	0.21
Reovirus (reo)[b] (Shatkin, Sipe and Loh, 1968)	1	2.5
	2	2.4
	3	2.3
	4	1.6
	5	1.6
	6	1.4
	7	0.92
	8	0.76
	9	0.64
	10	0.61

[a] These viruses are generally considered to contain only eight authentic genomic modules

[b] These RNAs are doubled-stranded

Now for fixed L, an increase in n will proportionately decrease the average size of the RNA modules. If this holds for RNA B as well, we may write $L_B \approx fL/n$, so that

$$K = 1 + g(\sigma)(1-q)L(1 - f/n) \qquad (13)$$

where f is a constant >1.

Hence, the important conclusion emerges that the greater n is, the greater K becomes (Fig. 2). In reality, various counteracting factors would come into play at sufficiently large values of n. There could, for example, be some variation of the mean σ with n due to conformational co-operativity, which would be expected to cause K to increase less rapidly for larger n. In any case, we predict that n values in monocompartment viruses should range to much larger values than in multicompartment viruses. This is as observed (Table 2).

Discussion

One of the basic theorems of physics states that information cannot be transmitted over long pe-

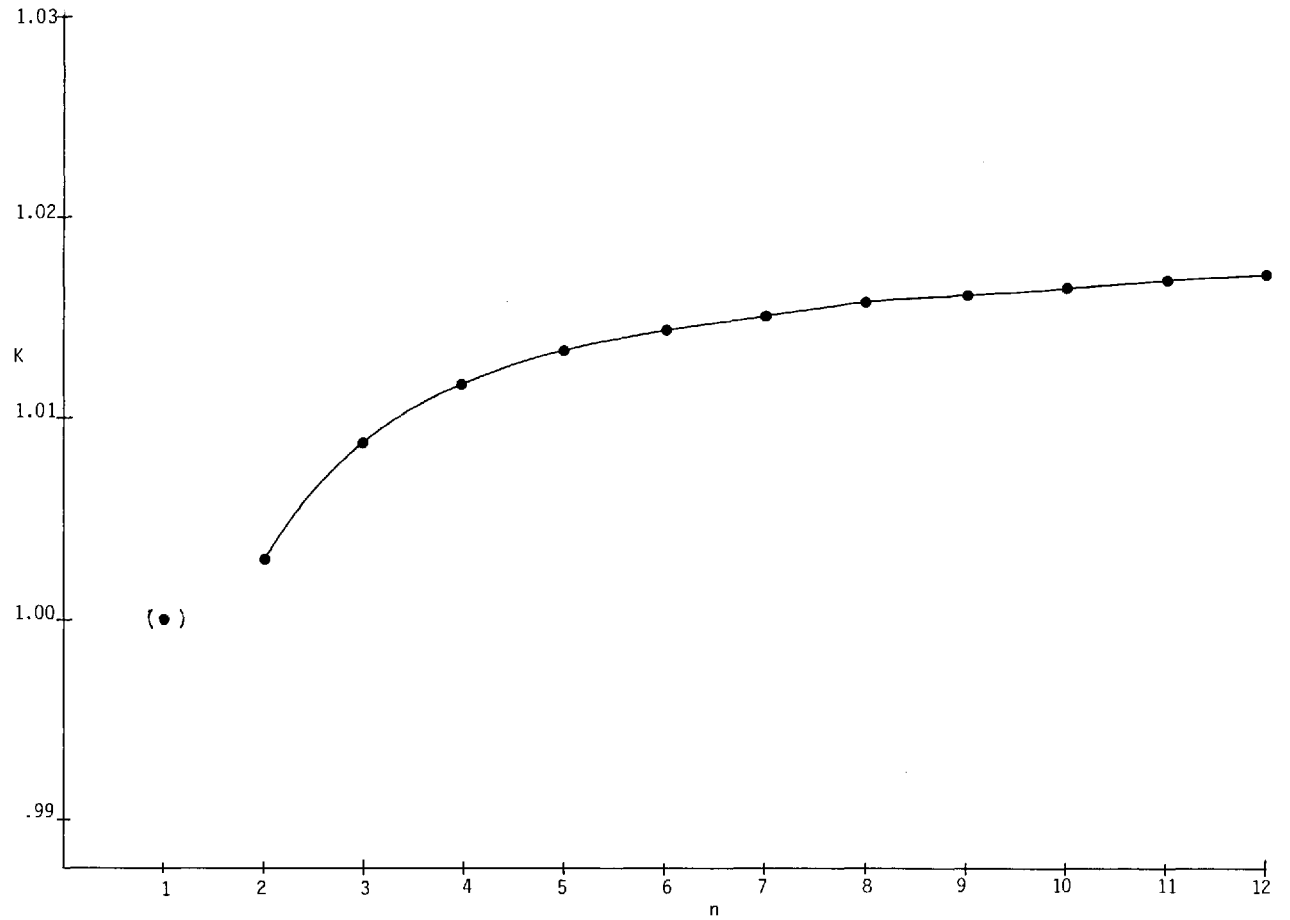

Fig. 2. Variation of relative advantage of divided genome strategy, K, with number of modules, n, for the monocompartment case [see Eq. (13)]. $g(\sigma) = 0.2$; $f = 1.7$; $(1-q)L = 0.10$ (10% errors)

riods of time without experiencing some deterioration in quality due to 'noise' in the copier process (see Shannon 1949). The error rate in the 'first' genes in prebiotic systems can be determined by measuring the rate at which non-complementary bases are incorporated into single-stranded RNA in the absence of enzymes. This value is about 10^{-1}–10^{-2} substitutions per nucleotide per doubling (Inoue and Orgel 1983). By contrast, in a present-day 'high-fidelity' system based on duplex DNA, the error rate can be as low as 10^{-11} (Drake 1974). Thus, during the course of evolution, noise levels in genetic replication have dropped by a factor of 100 million or so (Reanney 1984). This reduction has been brought about by the development of such corrective mechanisms as proofreading and mismatch repair.

RNA cannot use any of these restorative processes, so RNA viral genomes are the only (surviving?) genetic systems in which the error rate remains at the level (3×10^{-4}) characteristic of unrepaired, enzymatically catalysed nucleic acid synthesis. This presents RNA genomes with a set of unique survival problems (Reanney 1982) and limits the amount of information that can be encoded in any single uninterrupted RNA molecule to about 23,000 nucleotides (Eigen and Schuster 1977; Reanney 1984). The premise of this article is that genome segmentation compensates for this high error level by dividing the genetic data among smaller subunits, thus presenting a lesser target size to the various error-promoting agents. The protection offered by segmentation thus applies not only to the deleterious effects of error-prone replicases (the chief source of error) but also to damage by physical agents such as heat.

Most metabolically active cells contain large numbers of RNase enzymes to maintain a rapid rate of mRNA turnover. Since a single endonucleolytic cut normally destroys infectivity, RNases pose a fundamental problem for the intracellular survival of RNA genomes. Most single-strand RNAs appear to minimise this danger by folding into compact, largely double-helical formats [e.g., the flower arrangement for the coat protein gene of ribophage MS2 (Min Jou et al. 1972)] that are relatively resistant to the action of most RNases, whose preference is for single-stranded sites. It is possible that the fully base-paired, double-helical character of segmental reoviral RNAs is a consequence of the need to protect the large amount of information in the reoviral genome from enzymatic degradation.

One must remember here that duplex RNA is not functionally equivalent to duplex DNA, since the latter can be unwound, whereas the former cannot (Reanney 1982).

RNases may also favour evolution towards small segmental genomes because the probability of a single cut being introduced into an RNA molecule should increase monotonically with length. (In fact, one can show mathematically from a simple Brownian-motion model that the RNA degradation rate should vary with RNA length as $L^{1/6}$.) It may also be true that the chemical lability of RNA molecules at physiological pH favours genome segmentation, since the greater the number of phosphodiester bonds (i.e., the longer the molecule), the greater the chance of hydrolytic self-destruction (Reanney 1984). Thus, information divided between two (or more) RNA modules may have a significantly longer half-life in an RNase-rich, alkaline environment than would a continuous RNA of equivalent size.

The model developed in this paper shows that segmentation has followed two evolutionary routes which appear to be quite unrelated. Where modular RNAs are partitioned among discrete particles (multicompartment viruses) segmentation per se has a significant protective effect provided (a) the transmission of particles remains independent and random and (b) effective mechanisms exist for the host-to-host passage of viral genes. Point (a) is certainly true of natural viruses and point (b) is probably true of viruses that have insect vectors or that infect hosts (e.g. plants) that are grouped closely together in a common habitat (Nahmias and Reanney 1977). However, this kind of arrangement has the disadvantage that two or more particles must jointly infect a common cell if infection is to be successful. This limits the number of modular RNAs to two or three (Table 1 and Fig. 2). The sizes of the modular RNAs are also curtailed by this requirement, as the chances of successful coinfection are proportional to the number of available progeny viruses, which is in turn a function of genome lengths (the smaller the composite genome, the larger the number of RNAs that can be generated from a fixed pool of precursor elements). Thus, it is not surprising that multicompartment viruses in general have relatively small aggregate genome sizes. The problems associated with a multihit infection process may also explain why many RNA viruses have *not* adopted the divided genome strategy.

The second evolutionary route was followed by those viruses whose modular RNAs were able to combine in a sequence-specific manner. Such specific interactions, if spread over a large enough number of nucleotides and/or stages, provide for some degree of discrimination against miscopied information. The improved overall fidelity that can be achieved by this mechanism allows the amount of genetic information that can be encoded in RNA genomes to expand significantly. Only a few groups of RNA viruses display this feature. Chief among these are the reoviruses, which have both the largest and the most highly divided of all RNA genomes.

Is the model advanced in this paper sufficient to explain the abundance of divided genomes in RNA viruses? Lane (1979) lists four possible advantages of genome segmentation: (a) increased genetic flexibility; (b) more efficient packaging; (c) more efficient control of translation; and (d) increased resistance to inactivation by such environmental agents as ultraviolet radiation (UV).

The argument for increased flexibility is essentially that reassortment combines genetic information from different sources and so provides an RNA version of the 'hybrid vigour' seen in higher systems. However, the idea that divided genomes have been selected for because they *enhance* genetic variation (see Joklik 1974) seems paradoxical in a situation in which the amount of inherent genetic variation is so great that the genome can only be defined in a probabilistic sense (Domingo et al. 1978; Reanney 1982). Thus, while there is no doubt that reassortment among modular RNAs occurs in nature, the 'expanded variation' model seems inadequate to explain the genesis and maintenance of divided genomes among so many groups of RNA viruses.

Point (b) (more efficient packaging) is suspect because it would apply to single-stranded DNA as well as single-stranded RNA, and any credible theory of the origin of divided genomes must explain why these structures are virtually confined to RNA (as opposed to DNA) viruses. Point (d), increased resistance to UV, supports the general argument of this paper, since UV-induced lesions in DNA viruses can be repaired, whereas those in RNA viruses cannot.

This leaves point (c), more effective control of translation. Eukaryotic cells, unlike prokaryotes, transcribe their genes into monocistronic messenger RNAs. Jaspers (1974) has suggested that the replicative strategies of most groups of RNA viruses can be rationalised by assuming that they represent attempts to accommodate polycistronic RNAs to a biochemical environment tailored to process only monocistronic mRNAs (for a discussion see Reanney 1982). On this basis segmentation has the striking advantage that it divides RNA viral information into small units that closely resemble cellular mRNAs. There is no doubt, in our view, that this argument is correct, as far as it goes. However, it cannot be put forward as a general or unitary explanation of the divided genome phenomenon because other strategies open to and adopted by RNA viruses also enable RNA genomes to survive in na-

ture. Thus, some RNA viruses, e.g., the polio group, overcome the problem of translation by translating their polycistronic genomes into long polyproteins which are then cleaved into specific, functional peptides. Other viruses, e.g., the influenza group, generate monocistronic RNAs by transcription from continuous, negative-strand genomes using discrete initiation and termination signals. Yet another mechanism has been adopted by the coronaviruses, and so on (for a discussion of these various points, see Reanney 1982).

But perhaps the most telling argument against the 'monocistronic message' concept as a general explanation for RNA genome segmentation is the presence of RNA viruses with divided genomes in prokaryotes, since, as stated, prokaryote messengers are polycistronic, not monocistronic. In the context of a bacterial cell the segmental RNA genomes of the cystoviruses (see Matthews 1979) seem very much out of place, suggesting that, at least in this instance, a different explanation of the divided genome phenomenon must be sought.

In summary, we believe the various selective advantages of genome segmentation proposed to date fail, either singly or together, to take account of what we believe to be the chief guiding influence on the evolution of RNA as opposed to DNA viruses, namely the 10^5–10^8-fold greater error rate of RNA replication compared with DNA replication. Although factors such as the need to adapt eukaryotic RNA viral genomes to the unit-message character of higher cells may have played a part in the tendency of RNA genomes to split into separate modules, any explanation that does not recognise the critical importance of genetic noise for the divided genome phenomenon is at best only a partial answer, and at worst, a misleading one. We suggest that the current theories be revised to accommodate the model presented in this paper.

Appendix 1

We seek the values of L_i that will maximize the function

$$K = \sum_i^n N_i L_i q^{L_i} \prod_{j \neq i}^n \lambda_j \Big/ q^L \sum_i^n N_i L_i \tag{1}$$

subject to

$$\sum_i^n L_i = L \tag{2}$$

A simple expression may be obtained in the following manner. Consider a change in K caused by an increase of L_j by a small amount δ and a corresponding decrease of L_k by δ. The condition (2) remains valid, and for the L_i values giving a maximal value of K, $\delta K = 0$. That is, for $K = K(L_1, L_2, \ldots, L_n)$,

$$\delta K = 0 = K(L_1, L_2, \ldots, L_{j+}\delta, \ldots, L_{k-}\delta, \ldots, L_n) - K(L_1, L_2, \ldots, L_j, \ldots, L_k, \ldots, L_n) \tag{3}$$

Keeping only first order terms in δ, and using the excellent approximation $q^{L+\delta} = q^L[1 - \delta(1 - q)]$, substitution yields

$$q^{L_j} N_j [1 - L_j(1 - q)] \prod_{\nu \neq j} \lambda_\nu - q^{L_k} N_k [1 - L_k(1 - q)] \prod_{\nu \neq k} \lambda_\nu = (N_j - N_k) \frac{\sum_i N_i L_i q^{L_i} \prod_{\nu \neq i} \lambda_\nu}{\sum_i N_i L_i} \tag{4}$$

valid for all j,k.

This gives a rather complex set of relations between the L_i, N_i and q. An additional set may be obtained by setting $\delta K = 0$ for variations in the N_i. However, the biology of the situation allows an appropriate simplification. Any real virus system must cope with a range of values of q, since daily or seasonal temperature changes may affect q considerably, as may factors such as changing levels of ultraviolet radiation. The most generally valid solution to Eq. (4) will then be one that treats q as an independent variable and sets all individual coefficients of q^{L_i} identically equal to zero, since the q^{L_i} are independent functions (unless the L_i are equal, which then gives our final result immediately).

The only terms of the form q^{L_i} for $i \neq j,k$ come from the right-hand side of Eq. (4), and their coefficients can only be zero if $N_j = N_k$, i.e., $N_i = N/n$ for all i. In this case all λ_i are equal, and there is no solution for the coefficients of q^{L_j} and q^{L_k} unless $L_j = L_k = L/n$, as the reader may readily verify.

Appendix 2

We present here some details of a rigorous model of the replication of monocomportment divided genome viruses. The basic kinetic equations for the encapsidation reactions $A + B \rightarrow AB$, $A^* + B \rightarrow A^*B$, $A + B^* \rightarrow AB^*$ and $A^* + B^* \rightarrow A^*B^*$ may be written as

$$\frac{d[AB]}{dt} = k[A][B] \tag{1a}$$

$$\frac{d[AB^*]}{dt} = k\sigma_B[A][B^*] \tag{1b}$$

$$\frac{d[A^*B]}{dt} = k\sigma_A[A^*][B] \tag{1c}$$

$$\frac{d[A^*B^*]}{dt} = k\sigma_A\sigma_B[A^*][B^*] \tag{1d}$$

where k is the basic reaction constant and the σs indicate mean reductions of reaction rate for copies with errors. We have assumed second-order kinetics.

In addition, there are four equations of material conservation:

$$[A] + [AB^*] + [AB] = [A]_o \tag{2a}$$

$$[B] + [A^*B] + [AB] = [B]_o \tag{2b}$$

$$[A^*] + [A^*B] + [A^*B^*] = [A^*]_o \tag{2c}$$

$$[B^*] + [AB^*] + [A^*B^*] = [B^*]_o \tag{2d}$$

where $[A]_o$ = concentration of A produced in the first generation (proportional to q^{L_A}) and $[A^*]_o$ = concentration of A^* produced in first generation (proportional to $1 - q^{L_A}$). Similar definitions apply for $[B]_o$ and $[B^*]_o$.

It is possible to provide a straightforward interpretation of the σs as follows. The rate constant k of Eq. (1a) may be assumed to follow an Arrhenius-type equation

$$k = Ce^{-E/kT} \tag{3}$$

where C is a constant and E is the activation energy of the reaction $A + B \rightarrow AB$. Consider now the reaction $A + B^* \rightarrow AB^*$. Let the presence of each error in B^* increase the activation energy by ϵ,

so that the activation energy for this reaction is $E + m_B\epsilon$, where m_B is the mean number of errors in B*. The reaction constant for this second reaction is then $Ce^{-(E+m_B\epsilon)/kT}$, which by comparison with Eq. (1b) yields

$$\sigma_B = e^{m_B\epsilon/kT} \tag{4}$$

and similarly for σ_A.

To solve Eqs. (1) and (2) we differentiate Eqs. (2a)–(2d) with respect to time and substitute in Eqs. (1a)–(1d) to obtain

$$\frac{-d[A]}{dt} = k[A]\{[B] + \sigma_B[B^*]\} \tag{5a}$$

$$\frac{-d[B]}{dt} = k[B]\{[A] + \sigma_A[A^*]\} \tag{5b}$$

$$\frac{-d[A^*]}{dt} = k\sigma_A[A^*]\{[B] + \sigma_B[B^*]\} \tag{5c}$$

$$\frac{-d[B^*]}{dt} = k\sigma_B[B^*]\{[A] + \sigma_A[A^*]\} \tag{5d}$$

Eqs. (5a) and (5c) may be combined and integrated to yield

$$\frac{[A^*]}{[A^*]_o} = \left(\frac{[A]}{[A]_o}\right)^{\sigma_A} \tag{6a}$$

and similarly Eqs. (5b) and (5d) yield

$$\frac{[B^*]}{[B^*]_o} = \left(\frac{[B]}{[B]_o}\right)^{\sigma_B} \tag{6b}$$

Further integration of Eqs. (5a) and (5b) yields

$$[A] + [A^*]_o\left(\frac{[A]}{[A]_o}\right)^{\sigma_A} - N_A = [B] + [B^*]_o\left(\frac{[B]}{[B]_o}\right)^{\sigma_B} - N_B \tag{7}$$

where $N_A = [A]_o + [A^*]_o$ and $N_B = [B]_o + [B^*]_o$

This equation is transcendental and permits no further integration without the approximation of a very high error rate, that is, $[A^*]_o/[A]/ \gg 1$ and $[B^*]_o/[B] \gg 1$. This unfortunately does not correspond to any known physical system. If for simplicity we also assume $N_A = N_B$, the formal solution for [AB] is

$$[AB] = \frac{[A]_o[B]_o}{[B^*]_o}\left(\frac{[B^*]_o}{[A^*]_o}\right)^{1/\sigma_A}\frac{1}{(\sigma_A + \sigma_B - \sigma_A\sigma_B)}$$

$$\cdot[1 + (1 + k[B_o^*]\sigma_A\sigma_B t)^{1-\frac{1}{\sigma_A}-\frac{1}{\sigma_B}}] \tag{8}$$

so that as $t \to \infty$, and dropping concentration brackets,

$$AB_\infty = \frac{A_oB_o}{B_o^*}\left(\frac{B_o^*}{A_o^*}\right)^{1/\sigma_A}\frac{1}{\sigma_A + \sigma_B - \sigma_A\sigma_B} \tag{9}$$

For $\sigma_A = \sigma_B = 1$, and since $A_o^* = B_o^*$ to the accuracy of the high error approximation, $AB_\infty \approx A_oB_o/N_A$, which is the statistically correct result. For $A_o^* = B_o^*$ and $\sigma_A \approx \sigma_B$, we may write

$$AB_\infty = \frac{A_oB_o}{N_A}\frac{1}{\sigma(2 - \sigma)} \tag{10}$$

so that

$$F = \frac{AB_\infty}{N_B} = \frac{q^L}{\sigma(2 - \sigma)} \tag{11}$$

This formula is readily generalized to an n-module system by the consideration of successive reactions $A + B \to AB$, $AB + C \to ABC$, $ABC + D \to ABCD$, . . . , with the result

$$F = \frac{q^L}{\sigma^n(2 - \sigma)^n} \tag{12}$$

so that the selective advantage $K = F/q^L$ is

$$K = \frac{1}{\sigma^n(2 - \sigma)^n} \tag{13}$$

a monotonically increasing function of n and monotonically decreasing function of σ. Steps in the approximate solution disallow the limit $\sigma \to 0$.

References

Ahmed R, Fields BN (1981) Reassortment of genome segments between reovirus defective interfering particles and infectious virus: construction of temperature sensitive and attenuated viruses by rescue of mutations from DI particles. Virology 111:351–363

Bancroft JB (1972) A virus made from parts of the genomes of brome mosaic and cowpea chlorotic mottle viruses. J Gen Virol 14:223–228

Both GW, Bellamy AR, Street JE, Siegman LJ (1982) A general strategy for cloning double-stranded RNA: nucleotide sequence of the Simian-II rotavirus gene. Nucleic Acids Res 10: 7075–7087

Bromley PA, Barry RD (1973) Characterisation of the RNA of fowl plague virus. Arch Gesamte Virusforsch 42:182–196

Domingo E, Sabo D, Taniguchi T, Weissman C (1978) Nucleotide sequence heterogeneity of an RNA phage population. Cell 13:735–744

Drake JW (1974) The role of mutation in microbial evolution. Soc Gen Microbiol Symp (Cambridge) 24:41–58

Eigen M, Schuster P (1977) The hypercycle. A principle of natural self-organization. Part A: Emergence of the hypercele. Naturwissenschaften 64:541–565

Greenberg HB, Wyatt RG, Kapikian AZ, Kalica AR, Flores J, Jones R (1982) Rescue and serotypic characterisation of noncultivable human rotavirus by gene reassortment. Infect Immun 37:104–109

Haber S, Ikegami M, Bajet NB, Goodman RM (1981) Evidence for a divided genome in bean golden mosaic virus, a gemini virus. Nature 289:324–326

Habili N, Francki RIB (1974) Comparative studies on tomato aspermy and cucumber mosaic viruses. III. Further studies on the relationship and construction of a virus from parts of the two viral genomes. Virology 61:443–449

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

Inoue T, Orgel LE (1983) A non-enzymatic RNA polymerase model. Science 219:859–862

Jaspers EMJ (1974) Plant viruses with a multipartite genome. Adv Virus Res 19:37–149

Joklik W (1974) Evolution in viruses. Soc Gen Microbiol (Cambridge) 42:293–320

Joklik W (1981) Structure and function of the reovirus genome. Microbiol Rev 45:483–501

Kornberg A (1980) DNA replication. WH Freeman and Co, San Francisco, p 724

Lane LC (1979) The RNAs of multipartite and satellite viruses of plants. In: Hall TC, Davies JW (eds) Nucleic acids in plants, vol 2. CRC Press, Boca Raton, pp 65–110

Loeb AA, Kunkel TA (1982) Fidelity of DNA synthesis. Annu Rev Biochem 51:429–457

Matthews REF (1979) Classification and nomenclature of viruses. Intervirology 12:129–296

Maynard-Smith J (1978) The evolution of sex. Cambridge University Press, Cambridge, England, chapter I

Min Jou W, Haegeman G, Ysebaert M, Fiers W (1972) Nucleotide sequences of the gene coding for the bacteriophage MS2 coat protein. Nature 237:82–88

Nahmias AJ, Reanney DC (1977) The evolution of viruses. Annu Rev Ecol Systematics 8:29–49

Palese, P, Schulman JL (1976) Differences in RNA patterns of influenza A viruses. J Virol 17:876–884

Palese P, Young JF (1982) Variation of influenza A, B and C viruses. Science 215:1468–1473

Portner A, Webster RG, Bean WJ (1980) Similar frequencies of antigenic variants in Sendai, vesicular stomatitis and influenza A viruses. Virology 104:235–238

Prabhakar BS, Haspel MV, McClintock PR, Notkins AL (1982). High frequency of antigenic variants among naturally occurring human Coxsackie B4 virus isolates identified by monoclonal antibodies. Nature 300:374–376

Reanney DC (1982) The evolution of RNA viruses. Annu Rev Microbiol 36:47–73

Reanney DC, Pressing J (1983) Heat as a determinative factor in the evolution of genetic systems. J Mol Evol, submitted

Reanney DC (1984) Genetic noise in evolution? Nature 307: 318–319.

Reanney DC (1984) The molecular evolution of RNA viruses. Soc Gen Microbiol Symp (Cambridge) 35:175–196

Reijnders L (1978) The origin of multicomponent small ribonucleoprotein viruses. Adv Virus Res 23:79–102

Rogers J, Wall R (1980) A mechanism for RNA splicing. Proc Natl Acad Sci USA 77:1877–1879

Rose M, Doolittle WF (1983) Parasitic DNA—the origin of species and sex. New Scientist 16:787–789

Shannon CE (1949) The mathematical theory of communication. In: Shannon CE, Weaver W (eds) The mathematical theory of communication. University of Illinois Press, Urbana, Illinois

Shatkin AJ, Sipe JD, Loh P (1968) Separation of ten reovirus genome segments by polyacrylamide gel electrophoresis. J Virol 2:986–991

Silverstein SC, Christman JK, Acs G (1976) The reovirus replicative cycle. Annu Rev Biochem 45:375–408

Tinoco I, Uhlenbeck O, Levine M (1971) Estimation of secondary structure in ribonucleic acids. Nature 230:362–367

Webster RB, Granoff A (1974) The evolution of orthomyxoviruses. In: Kurstak E, Maramorosch K (eds) Viruses, evolution and cancer. Academic Press, New York, pp 625–647

Received July 18, 1983/Revised November 20, 1983