Error catastrophe for viruses infecting cells: analysis of the phase transition in terms of error classes

BY JULIA ALONSO1 AND HUGO FORT2,*

1Instituto de Física, Facultad de Ingeniería, Universidad de la República, Julio Herrera y Reissig 565, 11300 Montevideo, Uruguay
2Instituto de Física, Facultad de Ciencias, Universidad de la República, Iguá 4225, 11400 Montevideo, Uruguay

RNA viruses offer a very exciting arena in which to study evolution in ‘real time’ owing to both their high replication rate—many generations per day are possible—and their high mutation rate, leading to a large phenotypic variety. They can be regarded as a swarm of genetically related mutants around a dominant or master genetic sequence. This system is called a ‘viral quasi-species’. Thus, a common framework to describe RNA viral dynamics is by means of the quasi-species equation (QSE). The QSE is in fact a system of a very large number of nonlinear coupled equations. Here, we consider a simpler formulation in terms of ‘error classes’, which groups all the sequences differing from the master sequence by the same number of genomic differences into one population class. From this, based on the analogies with Bose condensation, we use thermodynamic inspired observables to analyse and characterize the ‘phase transition’ through the so-called ‘RNA virus error catastrophe’.

Keywords: viral quasi-species; error classes; error catastrophe; finite-size scaling

1. Introduction

RNA viruses have become especially popular for experimental evolution, owing to their rapid replication rates (Domingo & Holland 1994), which allow studies to run for hundreds of generations (Burch & Chao 2000). Examples of RNA viruses are HIV, influenza A (H1N1), foot and mouth disease, hepatitis C and poliovirus. In addition, the high mutation rates of RNA viruses generate a large genetic variation. Thus, the quasi-species (QS) concept introduced by Eigen (1971) is helpful to address the process of the Darwinian evolution of RNA viruses (Domingo 1992; Eigen & Biebricher 1998). A QS refers to the equilibrium spectrum of closely related mutants, dominated by a master sequence, that is generated by a mutation–selection process (Domingo 2006).

*Author for correspondence (hugo@fisica.edu.uy).

One contribution of 13 to a Theme Issue ‘Complex dynamics of life at different scales: from genomic to global environmental issues’.
On the one hand, the heterogeneous structure of the population can act as a reservoir of phenotypical variants potentially useful against environmental changes. On the other, there appear to be limits on how advantageous a high mutation rate can be, and there is evidence that, if an RNA virus QS goes beyond a mutation threshold, the population will no longer be viable. The phenomenon that occurs when the loss of genetic fidelity results in a lethal accumulation of errors has been termed the error catastrophe (Domingo & Holland 1994; Eigen & Biebricher 1998). This originated the controversy of survival of the fittest versus survival of the flattest (Wilke et al. 2001): a fast-replicating organism that occupies a narrow peak versus one that occupies a lower but flatter peak.

Here, we propose a framework to formulate problems like the above-mentioned trade-off between mutation and fitness and the error catastrophe. We try to maintain a balance between realism and simplicity. Unlike many models found in the literature, where viruses are modelled as free particles without any interaction with a host, here we explicitly consider viruses to be parasite binding organisms. In virology, the number of mutations per genome and replication round is often referred to and interpreted as an error class in the mathematical sense. We use this concept in our model, which greatly reduces the number of equations to be solved. Moreover, among the many possible distributions of fitness among the error classes, we choose the simplest one or Eigen landscape. This consists in assuming that all the classes have the same fitness except for a single master class that has a higher fitness (see §3).

The use of physical tools for the analysis of the dynamics of various biological systems is becoming increasingly common. RNA viruses in particular represent a very exciting challenge in this regard because (i) they exhibit an ‘order parameter’ that becomes zero at the error catastrophe, and (ii) the diversity of a QS can be straightforwardly quantified by means of the entropy. These two facts suggest an analogy with the phenomenon of Bose condensation. This, in turn, allows us to analyse the critical behaviour of thermodynamic-like observables like an order parameter and specific heat (see §4), and to characterize the phase transition of the system (see §5).

2. From the molecular quasi-species model to the quasi-species basic model of viral dynamics

QS theory was proposed by Eigen (1971) to describe the self-replication of biological macromolecules with the aim of understanding the origin of life. Since then, the QS concept has been applied to RNA virus populations, reflecting their large genetic variability (Domingo et al. 1985). For some time, some researchers regarded this approach as incompatible with the more conventional approach of population dynamics, namely population genetics. However, later it was realized that they are simply two equivalent approaches to describe the evolution (mutation–selection balance) of RNA viruses (see Wilke 2005): while population genetics involves a top-down approach to the problem, QS theory, based on chemical kinetics from its birth, corresponds to a bottom-up approach.

QS theory enables virologists to describe Darwinian evolution even when RNA replication occurs in a cell-free system. Given a population of organisms, with $n$
different types of sequences, the genomic structure of the population is given by the vector:

\[ \mathbf{v} = (v_1, v_2, \ldots, v_n), \] (2.1)

where \( v_i \) is the population of those organisms whose genome sequence is of type \( i \).

The quasi-species equation (QSE), describing the selection–mutation balance among sequences and then governing the evolution of \( \mathbf{v} \), is

\[ \frac{d\mathbf{v}}{dt} = W\mathbf{v} - \bar{A}(\mathbf{v})\mathbf{v}, \] (2.2)

where \( W \) is the selection–mutation matrix whose elements are \( W_{ij} = A_j Q_{ij} \), \( A_j \) is the replication rate of sequence \( j \), \( Q_{ij} \) is the probability that a \( j \) sequence mutates to an \( i \) sequence and \( \bar{A}(\mathbf{v}) \) is the mean replication rate of the population. The eigenvector associated with the largest eigenvalue of \( W \) is the steady-state solution of the previous equation and is known as the QS.

The population of a given variant, \( v_i \), depends not only on its intrinsic replication rate \( A_i \) but also on the rate of replication of the other variants (which may occur due to errors in the replication of these) and their populations. As a consequence, the target of the evolutionary process (selection–mutation) is not the individual sequence \( i \), with its intrinsic or own replication rate \( A_i \), but the QS as a whole entity. What matters is the effective replication rate of a particular genome \( i \), which is the fitness of that genome. The set of fitness values for sequences that comprise the QS is known as the fitness landscape. The most common fitness landscape is the Eigen landscape or single-peak landscape (mentioned in §1): a single fitness peak at the master sequence emerging from a plateau of constant height. In this context, the QSE model predicts an error threshold (critical mutation rate) above which the master sequence is extinguished. To make this point more clear, it is helpful to reduce the vector equation (2.2) to only two scalar equations by a mean-field approximation: one corresponding to the master sequence’s population \( v_m \) (defined as the most frequent in the population) and the other describing the mutant cloud composed by all the remaining sequences. Let \( A_m \) be the replication rate of the master sequence and \( A_c \) the replication rate of the cloud. A common additional simplification is to neglect the small probability of mutations from the cloud to the master sequence (Nowak 1992). Under this simplifying assumption, if \( M \) is the mutation rate of the master sequence, the fractional population of the master sequence becomes

\[ v_m^{\text{MF}} = \begin{cases} 1 - \frac{M}{M_c}, & M < M_c, \\ 0, & M > M_c, \end{cases} \] (2.3)

where \( M_c = 1 - A_c/A_m \) and the superscript MF is to remind us of the mean-field approximations we are considering. Figure 1 shows the behaviour of \( v_m^{\text{MF}} \) as a function of \( M/M_c \) as given by equations (2.3).

Nevertheless, a virus is an obligatory intracellular parasite, which means that it needs to enter into a cell to use its functions, and this fundamental biological aspect is not taken into account in the QSE. A simple model that considers cells, and distinguishes between infected and non-infected cells, is the so-called basic...
model of viral dynamics (BMVD), proposed by Nowak & May (2000). They also extended the model to include the concept of QS in it, yielding a QS-BMVD given by the following equations:

\[
\begin{align*}
\frac{dx}{dt} &= \lambda - d x - x \sum_i \beta_i v_i, \\
\frac{dy_i}{dt} &= x \sum_j Q_{ij} \beta_j v_j - a_i y_i, \quad (2.4) \\
\frac{dv_i}{dt} &= k_i y_i - u_i v_i,
\end{align*}
\]

where the time-dependent variables in this model are: the population of uninfected cells \(x\), the population of free viral particles with the \(i\)th viral genome or sequence and \(y_i\) the population of infected cells that produce the \(i\)th sequence. The dynamics of infection is described by chemical kinetics using the following coefficients/terms:

- \(\beta_i\) is the infectivity of virus particles of type \(i\),
- \(x\beta_i v_i\) is the rate at which uninfected cells react with viral particles of type \(i\),
- \(\lambda\) is the immigration rate,
- \(d x\) is the rate of normal mortality (in the absence of virus),
- \(k_i y_i\) is the rate at which infected cells produce viral particles,
- \(u_i v_i\) is the rate at which viral particles die,
- \(a_i y_i\) is the rate at which infected cells die,
- \(x Q_{ij} \beta_j v_j\) represents the mutational gain of type \(i\): when \(j = i\) then the \(j\)-genome (an \(i\)-genome actually) is replicated without error in the cell, while if \(j \neq i\) the \(j\)-genome is replicated with error, becoming a genome of type \(i\).

Notice that the equation for virus evolution is different from QSE (2.2).
This model contains parameters that are not easy to determine. To overcome this obstacle, we rewrite it in a dimensionless form. We assume the same intrinsic rate of death for all infected cells, $a_i = a$, and then we measure time in units of $1/a$. So we rename $t/(1/a)$ as $t$, $x/(\lambda/d)$ as $x$, $y_i/(u_i d/k_i \beta_i)$ as $y_i$ and $v_i/(d/\beta_i)$ as $v_i$. In this way, equations (2.4) can be rewritten as

$$
\frac{d}{a} \left( 1 - x - \sum_i v_i \right),
$$

$$
\frac{dy_i}{dt} = R_i x \sum_j Q_{ij} v_j - y_i,
$$

and

$$
\frac{dv_i}{dt} = \frac{u_i}{a} (y_i - v_i).
$$

At equilibrium it turns out that $y_i = v_i$, so we obtain the following equations for the equilibrium points:

$$
\frac{1}{1 + \sum_i v_i}
$$

$$
R_i \sum_j Q_{ij} v_j - \left( 1 + \sum_k v_k \right) v_i = 0.
$$

The free parameters left—besides mutation rates—are known as the basic reproductive ratios $R_i = (\lambda \beta_i k_i)/(d au_i)$ associated with each genome. The set of different reproductive ratios of each sequence takes the role of a fitness landscape in our model. Incidentally, when analyzing RNA viruses, virologists often consider the number of mutations in a genome with respect to another taken as a reference, e.g. the master sequence (see Crotty et al. 2001; Vignuzzi et al. 2006).

3. Viral dynamics in error classes

For the sake of simplicity, we consider binary sequences (bit sequences) so that the number of errors that lead to the transformation of one sequence into another is equal to the Hamming distance $d_H$ between these two sequences. Swetina & Schuster (1982) were the pioneers in taking a reference sequence (e.g. the master sequence) and putting together all sequences with the same Hamming distance to the reference sequence into what is mathematically known as an error class. Following this recipe, we reformulate the QS-BMVD model of Nowak & May (2000) in terms of error classes. This reduces the number of resulting equations and makes it easier to carry out numerical simulations: in bit sequences of length $L$ this number is reduced from $2^L$ to $L + 1$.

Let us denote by $\Gamma_K$ the set of different elements that make up the error class $K$ ($d_H = K$) and by $V_K$ its population. Then

$$
V_K = \sum_{i \in \Gamma_K} v_i.
$$
Assuming that all sequences in the same error class \( K \) have the same basic reproductive ratio \( R_K \), we obtain instead of equations (2.6)

\[
x^* = \frac{1}{1 + \sum_{J=0}^{L} V_J^*}
\]

and

\[
R_K \sum_{J=0}^{L} Q_{KJ}^* V_J^* - \left(1 + \sum_{J=0}^{L} V_J^*\right) V_K^* = 0.
\]

Here, the elements of mutation matrix \( Q' \) for error classes are given by

\[
Q'_{KJ} = \min\{K, J\} \sum_{P=\max\{K+J-L, 0\}}^{\min\{K, J\}} \left(\frac{J}{P}\right) \left(\frac{L-J}{K-P}\right) (1-\mu_b)^{L-P} \left(\frac{\mu_b}{1-\mu_b}\right)^{K+J-2P},
\]

where \( \mu_b \) is the single-bit error rate. Let us consider the Eigen single-peak landscape, where the master class \( K = 0 \) has the higher fitness, \( R_0 \), and all the other classes have a lower fitness, \( R = R_0/\sigma \), where \( \sigma \) is known, in the jargon of virology, as the superiority parameter of the master sequence \( (\sigma > 1) \). The previously considered MF approximation reduces the problem even further to just two different error classes, \( K = 0 \) and \( K = C \), for all the sequences in the cloud of mutants. In addition, if, as we did before, mutations from the class \( C \) to the class 0 are neglected, then one gets:

\[
Q'_{0C} = 0 \quad \text{(no mutations from cloud to master class)},
\]

\[
Q'_{CC} = 1 \quad \text{(mutations from cloud remain in the cloud)}
\]

and

\[
Q'_{C0} = (1-\mu_b)^L = 1-M \quad \text{(mutation rate for master class is } M)\).
\]

Substituting the above relations into equations (3.1) yields a simple expression for the fractional population of the master class (class 0), depending on whether \( M \) is above or below an error threshold \( M_c = 1 - \sigma^{-1} \):

\[
V_{0}^{MF} = \begin{cases} \frac{M_c - M}{(1-M)M_c}, & M < M_c, \\ 0, & M > M_c. \end{cases}
\]

Figure 2 shows the behaviour of the fractional population of the master class \( V_0^{MF} \) versus \( M/M_c \) for \( \sigma = 10 \) and the fractional population of the cloud \( (1-V_0^{MF}) \).

In figure 3, we show the corresponding ‘phase diagram’. The threshold curve separates the QS phase from the drift phase (Solé et al. 2006) in which the virus is no longer viable.

4. Observables

Clearly, the phenomenology of the transition experienced by viral populations from the QS to the drift phase is reminiscent of Bose condensation or the phase transition in the Landau two-fluid model describing the behaviour of superfluid helium or He II. Below the critical temperature \( T_c \) there coexist two components: a normal fluid and a superfluid component. As the temperature is decreased, more
Error catastrophe for viruses in cells

Figure 2. The fractional populations of the master class $V_{0}^{MF}$ (full line) and of the cloud (dashed line) as a function of $M/M_c$, for $\sigma = 10$. The resemblance to the Bose condensation is clear.

Figure 3. Mean-field phase diagram: the QS (drift) phase is below (above) the curve $1-\frac{M}{M_c}$. Solid line, $M_c = 1 - \frac{\sigma}{\sigma + 1}$.

and more normal fluid is transformed into the Bose condensate or superfluid. In the absolute zero temperature limit, 100 per cent of the system is condensate in He II. Comparing figure 2 with a graph of the superfluid and normal fractions of helium atoms, $n_s$ and $n_n$, versus the temperature $T$, the analogy is clear: the superfluid (normal) component corresponds to the master sequence (cloud of mutants) and the temperature with the mutation rate $M$. Based on this analogy, we take the fractional population of the master class, $V_0$, as the order parameter and $M$ plays a role analogous to temperature in physical systems.

Phil. Trans. R. Soc. A (2010)
Furthermore, this transition involves a change in the ‘biodiversity’ of the viral population, which is usually measured by its entropy (see, for example, Domingo 2007). Thus, from the usual definition of the entropy $S$, 

$$S = - \sum_{i=1}^{2^L} p_i \ln(p_i),$$  

(4.1)

where $p_i$ is the probability of finding a given sequence $i$ in the population of $2^L$ different types of sequences of viral QS. That is, $p_i$ coincides with the fractional population of the sequence $i$. If $i \in \Gamma_K$, then $v_i = V_K / C_K^L$, where $C_K^L$ is the degeneracy of the $K$ class, and equation (4.1) can be expressed as

$$S = - \sum_{K=0}^{L} V_K \ln \left( \frac{V_K}{C_K^L} \right).$$  

(4.2)

Numerical results for the entropy can be seen in figure 4 for $L = 100$ and $\sigma = 10$. Entropy increases as the presence of the master sequence decreases, i.e. it increases with $M$, which then plays the role of temperature. At the threshold $M_c$, the entropy reaches a maximum value $\ln(2^L - 1)$ and then remains constant for $M > M_c$. Following the analogy with thermal physics, a specific heat can be obtained from the entropy as

$$c = M \frac{dS}{dM}.$$  

(4.3)

It is worth mentioning that the analogy between the error threshold transition for QS and the phase transition taking place in other physical systems like the two-dimensional Ising model has been explored in many papers (see Leithäusser 1986, 1987; Tarazona 1992; Pastor-Satorras & Solé 2001). Additionally, this phase transition approach has been set in the context of cancer seen as a QS problem (see Solé 2003; Solé & Deisboeck 2004).
5. Critical exponents from finite-size scaling

To characterize a phase transition that is presumably continuous, it is necessary to analyse the singular behaviour that the relevant physical quantities of the system have at the critical point. These singularities take the form of power laws in the reduced temperature \( t = (T - T_c) / T_c \). The finite-size scaling (FSS) method is a straightforward way to extract the critical exponents by analysing how measured observables depend on changes in the size parameter \( L \) (see Newman & Barkema 1999). The following relations are only valid when the critical reduced \('temperature' or \('reduced mutation rate \( \epsilon = (M - M_c) / M_c \) is close to 0:

\[
\chi = L^{\gamma / \nu} \tilde{\chi}(L^{1/\nu} \epsilon),
\]

\[
c = L^{\alpha / \nu} \tilde{c}(L^{1/\nu} \epsilon)
\]

and

\[
m = L^{\beta / \nu} \tilde{m}(L^{1/\nu} \epsilon),
\]

where \( \nu, \alpha, \beta \) and \( \gamma \) are, respectively, the critical exponents for the correlation length \( \xi \), the specific heat \( c \), the order parameter \( m \) and the susceptibility \( \chi \). The scaling functions \( \tilde{\chi}, \tilde{c} \) and \( \tilde{m} \) in the above relations, by construction, should be independent of the size \( L \), but strongly vary with the parameters \( \epsilon, \nu, \alpha, \beta \) and \( \gamma \). If the correct values for the parameters are chosen, the data obtained for different system sizes will collapse onto a single curve. Before continuing, we want to make a remark concerning the critical exponent for the correlation length \( \nu \) appearing in equations (5.1)–(5.3). In our model, as in the combinatorial problems discussed by Kirkpatrick & Selman (1994) or the case of QSE studied by Campos & Fontanari (1998), there is no geometric criterion to define the correlation length \( \xi \). However, it is expected that the region where the transition occurs is reduced to 0 as \( L^{-1/\nu} \rightarrow 0 \) when \( L \rightarrow \infty \); then the exponent \( \nu \) can be associated with this characteristic behaviour. In this way, we can still use data collapsing to estimate the critical exponents \( \alpha \) and \( \beta \).

The results of numerical simulations for \( V_0 \) for \( L = 50, 70, 100, 120, 150 \) and 200 with \( \sigma = 10 \), together with the approximation by mean field \( V_0^{\text{MF}} \) (neglecting in the approximation mutations from the cloud to the master sequence) are shown in figure 5 in terms of \( M / M_c \). From this plot, it is possible to estimate the critical exponents \( \beta \) (corresponding to the behaviour of the order parameter in the vicinity of the transition) and \( \nu \).

As can be seen in figure 5, higher values of \( L \) give better similarity between \( V_0 \) and \( V_0^{\text{MF}} \). Hence, mean field is a good approximation in the thermodynamic limit \( (L \rightarrow \infty) \) and the critical temperature can be taken as \( M_c = 1 - 1/\sigma \). Near the transition, the order parameter \( V_0 \) is well approximated by \( V_0^{\text{MF}} \) and, therefore, we can assume that \( \beta = 1 \) as follows. Rewriting equation (3.3) for \( M \approx M_c \) \( (M < M_c) \) in terms of reduced temperature \( \epsilon = (M - M_c) / M_c \):

\[
V_0^{\text{MF}} \approx -\frac{1}{1 - M_c} \frac{M - M_c}{M_c} \propto -\epsilon^1.
\]

On the other hand, as seen in the inset of figure 5 for each \( L \), \( V_0 \) linearly approaches 0, vanishing for an \( M_L \) value that we can estimate by adjusting the values of \( V_0 \) by a straight line in the region where \( M / M_c \) is close to 1. Once we
have the points $M_L/M_c$ of intersection with the axis $M/M_c$ we can use a result from finite-size scaling:

$$\ln \left( \frac{M_L}{M_c} - 1 \right) = -\frac{1}{\nu} \ln(L) + \text{constant}. $$

Then, as we have $M_L/M_c$ values for each $L$, to obtain the slope ($-1/\nu$) it is enough to make a linear fit on the pairs of values $\ln(L)$ and $\ln(M_L/M_c - 1)$ as shown in figure 6. Finally, the exponent is:

$$\nu = 0.98 \pm 0.06. $$

Figure 7 shows that for $\nu = 0.98$, $\beta = 1$ and $M_c = 1 - 1/\sigma$, the collapse of data is very good.

The specific heat $c$ of the system can be obtained from the entropy by equation (4.3). Numerical results for $c$ for different values of $L$ are shown in figure 8. As we know from FSS, the maximum of the specific heat behaves as

$$c_{\text{max}} \sim L^{\alpha/\nu}. $$

Taking the logarithm on both sides, we obtain a straight line:

$$\ln(c_{\text{max}}) = \frac{\alpha}{\nu} \ln(L) + \text{constant}. $$

Then a linear fit on the pairs of values $\ln(L)$ and $\ln(c_{\text{max}})$ gives the slope of the line $\alpha/\nu$ (figure 9). As $\nu$ is known, $\alpha$ can be determined and the error in its estimate:

$$\alpha = 1.9 \pm 0.1. $$

Once we have determined $\alpha/\nu$, we can plot $\tilde{c}(\epsilon L^{1/\nu}) = c(\epsilon)L^{-\alpha/\nu}$ versus $\epsilon L^{1/\nu}$. Figure 10 shows that the data collapsing is quite good so the estimations for the
Error catastrophe for viruses in cells

Figure 6. Estimated exponent $\nu$: $\nu = 0.98 \pm 0.06$.

Figure 7. Collapse of the scaling functions $V_0 L^{\beta/\nu}$ versus the scaling variable $\epsilon L^{1/\nu}$. Open circles, $L = 50$; inverted triangles, $L = 70$; open squares, $L = 100$; open triangles, $L = 120$; crosses, $L = 150$; diamonds, $L = 200$.

Critical exponents $\beta$, $\nu$, $\alpha$ and the critical temperature are reliable. Summarizing we get

$$\begin{align*}
\nu &= 0.98 \pm 0.06, \\
\beta &= 1.0 \\
\alpha &= 1.9 \pm 0.1.
\end{align*}$$

Phil. Trans. R. Soc. A (2010)
Figure 8. Specific heat for different $L$ values versus $M/M_c$. Open circles, $L = 50$; inverted triangles, $L = 70$; open squares, $L = 100$; open triangles, $L = 120$; crosses, $L = 150$; diamonds, $L = 200$.

Figure 9. Estimated exponent $\alpha$: $\alpha/\nu = 1.97 \pm 0.01$.

6. Final comments

There are a large number of theoretical results based on analytic approaches to the full system of QSE that have led to relevant conclusions about many viruses. In this work, we attempt to simplify the insight into QS by taking into account a dimensionless form of the equations while including the host cells in which the viruses replicate. We have proposed a reformulation of the
Nowak & May (2000) QS-BMVD model, which incorporates cells, in terms of error classes. The utilization of error classes allows a great simplification of the system of differential equations to be solved. We use this new framework to analyse the error catastrophe phase transition. Exploiting analogies with widely studied models from physics, like, for example, Bose condensation, we identify a useful ‘order parameter’ to characterize the two phases separated by the error catastrophe. In addition, by taking the entropy as a diversity parameter for viral QS, we were able to calculate a ‘specific heat’ as another quantity to characterize the phase transition. As an application of statistical mechanics techniques, using the FSS method, we obtained the corresponding critical exponents.

Here, we considered the simplest Eigen landscape. In future work it should be interesting to explore what happens in the case of more realistic landscapes.

To conclude, we think that formulating QS in terms of physical observables provides new insights as well as alternative powerful quantitative techniques to approach RNA viruses.

H.F. thanks Juan Arbiza and Esteban Domingo for stimulating discussions and valuable comments. This work was supported in part by PEDECIBA and ANII (Uruguay).

References


