Role of Error Catastrophe in Transmission Ability of Virus

Naoyuki Takahata[1] and Hirotaka Sugawara[2]

[1] The Graduate University of Advanced Studies (Sokendai), Hayama, Kanagawa, Japan
[2] KEK, Tsukuba, Ibaraki, Japan

The role played by "error catastrophe" is explicitly taken into account in the mathematical formulation to analyze the COVID-19 data. The idea is to combine the mathematical genetics formalism of the error catastrophe of mutations in the virus gene loci with the standard model of epidemics which lacks the explicit incorporation of the mutation effect on the spreading of the viruses. We apply the formalism to the case of SARS-CoV-2 virus. We assume the "universality" of the error catastrophe in the process of analyzing the data. This means that some basic parameter to describe the error catastrophe is independent of which group (country or city) we deal with. Concretely, we analyze the omicron data of South Africa and then analyze the cases of Japan using the same value of the basic parameter derived in the South Africa analysis. The result of the excellent fittings of the two data, one from South Africa and the other from Japan with the common values of genetic parameters, justifies our universality assumption of these parameters.

Keywords: SARS-CoV-2, COVID-19, omicron, mutation, error catastrophe, APOBEC3, standard model of epidemiology

---

Correspondence should be addressed: N. Takahata, Shimoyamaguchi 120-3, Hayama, Kanagawa 240-0116, Japan (e-mail: takahata@soken.ac.jp) and H. Sugawara, KEK, 1-1 Oho, Tsukuba, Ibaraki 305-0801, Japan (e-mail: sugawara@post.kek.jp).

1. INTRODUCTION

In usual analysis of virus spreading based on the standard model of epidemiology by Kermack and McKendrick (1927), we ignore the possible role played by mutations, at least in the explicit form. One of the present authors published two papers (Sugawara, 2021a, b) analyzing the world-wide COVID-19 situation based on the standard model through which he learned that the standard model does not incorporate the role played by mutations properly. Because of this fact, the analysis encountered some difficulty in interpreting the result of the analysis.

A simplified assumption of the absence of mutations in the virus genome is still made even for recently developed models in which, for example, multiple pathogen strains interact to each other through cross-immunity (Kucharski, Andreasen and Gog, 2016 for a review) or strains may interfere through the removal of individuals from the susceptible pool after an acute infection (e.g., Rohani et al., 2002). Meanwhile, the importance of random nature of pathogen mutations and emergence of new strains has been well recognized and studied by simulation under the assumption that a host's risk of infection is proportional to the genetic distance between the infecting phenotype and the closest phenotype in the host's immune history (Bedford et al., 2012). Recently, Miller, Elenberg and Dubrawski (2022) explored an existing agent-based model for influenza A/H3N2 which incorporates mutation effects into both transmission rate and the antigenic space in the standard model.

We present here a way to remedy the current situation for the presence of largely deleterious but occasionally advantageous antigenic mutations by introducing into the scheme of the standard model some elements showing the vital role played by mutations before, during and after the period of extensive virus spreading. When we analyze the data of variant strains of SARS-CoV-2 such as α, β, γ, δ, omicron BA.1 and omicron BA.2, we notice that the time variations of the number of the infected show certain interesting similarity among the different variant strains. In this regard, we emphasize an outstanding role played by error catastrophe (Eigen, 1971): Viruses can improve their transmission ability by mutations. But it seems that there is a maximum ability they can reach due to the errors during the DNA/RNA replication processes. After they reach the maximum (error threshold), the forced mutations (by such an enzyme as APOBEC3 (Sadeghpour et al., 2021) as was recently proposed) keep causing errors in the

copying process of the virus multiplication and thus reducing the transmission power.

In this paper, we first explain the biological justification in more detail (section2). Then we formulate the process in mathematical terms (section 3). We apply the mathematical scheme developed in section 3 to actual data analysis taking the Tokyo case as an example (section 4). Section 5 is devoted to the analysis of omicron data from South Africa and Japan. The appendix deals with the issue of symbiosis in our formalism.

2. BIOLOGICAL JUSTIFICATION OF THE MODEL

As stated in the introduction, the analyses of the data of variant strains of SARS-CoV-2 show that the time variations of the number of the infected imply certain interesting similarity among the different variant strains. So far no clear explanation of this phenomenon exists within the framework of mathematical epidemiology. The data also show rather similar decaying patterns of the infection numbers of a certain variant once they reach the maximum in various countries where different epidemic policies are adopted including the vaccination rates. These similarities especially of the decaying patterns which are independent of the kind of variant strains or of epidemic policies seem to be explainable by assuming that the reason is on the side of viruses concerned.

A possible model is what we call an error catastrophe model (Eigen, 1971) which claims that the accumulated "error" in the mutation makes it hard for the multiplication process to be smoothly performed. The infection rate of such a virus as SARS-CoV-2 which has a rather high rate of mutations will be strongly influenced by the errors and its decaying patterns are largely determined by them. In the case of SARS-CoV-2, which allows APOBEC3 enzymes to directionally mutate from C to U, the error catastrophe may be particularly important. We do not need to treat this RNA editing by APOBEC3 distinctly from the usual mutation that occurs during the replication (Sadeghpour et al., 2021). It simply emphasizes the importance of "error catastrophe" and, therefore, it is hidden in our mathematical formulation.

The genome size of SARS-CoV-2 is about 30 kb, three times larger than that of typical RNA viruses such as influenza viruses and HIV. Despite being an RNA virus, CoV-2 accumulates mutations comparatively slowly under high fidelity of

replication by expressing a 3'-to-5' exoribonuclease nsp14-ExoN (Smith et al., 2013). However, if nsp14-ExoN is disrupted by self-mutagenesis or external agency, the mutation rate may increase 15 times (Robson et al., 2020; Thakur et al., 2022 for a review). It is observed that nearly one-half of nucleotide changes in CoV-2 are C to U transitions and occur preferentially in both 5'U/A and 3'U/A flanking sequence contexts that are favored motifs of the human APOBEC3 protein family of cytidine deaminases. As a result, CoV-2 may mutate at a rate of 3×10^{-4} /site/year, or 9/genome/year, consistent with a mean of 5.5 to 9.5 nucleotide differences between variants (Simmonds and Rampant, 2020; Ratcliff and Simmonds, 2021). Such high and context-dependent mutations likely hinder proper self-replication, lower the power of infection, though by yet unknown molecular mechanisms, and allow maladaptive amino acid changes to accumulate, leading to suicidal explosion of the viruses or error catastrophe. Even so, antigenic escape can occur and challenges pandemic control efforts. This unique paradigm of CoV-2 genome evolution is not incorporated by standard models of molecular epidemiology.

In the past, some efforts have been done to incorporate the pathogen evolution to the framework of epidemic models (SIR or its extension). For example, see section3 of Kucharski, Andreasen and Gog (2016). Compared to those works quoted there, our work is unique in its emphasis on the error catastrophe.

3. MATHEMATICAL FORMULATION

The purpose of this section, in a way, is to combine mathematical formulation of the mutation process with the standard model of mathematical epidemiology (Kermack and McKendrick, 1927; Kucharski, Andreasen and Gog, 2016). We start with a formulation of the mutation process and properly combine it with the standard model of epidemiology. Our target space is a well-defined population of humans such as a country or a city. We concentrate on the genes which are relevant for the transmission of the virus. This means that we concentrate mostly on the mutations in the spike proteins but not necessarily restricted to them.

We start by defining x_i to be the number of viruses that belong to the i-th variant. Here the "variant" means not just a prominent variant such as δ or omicron, but any group of viruses which has distinctive mutations from the original one. Let a_i denote the rate at which variant i reproduces (a_i does not

necessarily mean the infection rate of variant i), and let Q_{ij} denote the probability of a virus of variant i mutating to variant j. Then the rate of change of x_j is given by

$$\frac{dx_j}{dt} = \sum_i a_i Q_{ij} x_i \ .$$

Suppose we group one kind of variants which have larger infection rates into type X and those variants which have smaller infection rate into type Y. We respectively denote y_i, b_i and R_{ij} rather than x_i, a_i and Q_{ij} for the latter case. Then, we obtain

$$\sum_j \frac{dx_j}{dt} = \sum_{i,j} a_i (1 - Q_{ij}) x_i + \sum_{i,j} b_i R_{ij} y_j \ ,$$

and

$$\sum_j \frac{dy_j}{dt} = \sum_{i,j} b_i (1 - R_{ij}) y_i + \sum_{i,j} a_i Q_{ij} x_i \ .$$

To simplify the discussion, we assume a_i, b_i, Q_{ij}, and R_{ij} are all independent of suffix "i". Then we get

$$\frac{dx}{dt} = a(1 - Q)x + bRy \qquad [3.1]$$

and

$$\frac{dy}{dt} = b(1 - R)y + aQx. \qquad [3.2]$$

Here we have

$$x = \sum_j x_j, \ y = \sum_j y_j, \ Q = \sum_j Q_{ij}, \ R = \sum_j R_{ij}.$$

In the two-component vector form, Eqs. [3.1] and [3.2] become

$$\frac{d}{dt}\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} a(1-Q) & bR \\ aQ & b(1-R) \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} \equiv \Omega \begin{pmatrix} x \\ y \end{pmatrix}. \qquad [3.3]$$

Here, $1+a$ is the average reproduction rate of variant in type X and $1+b$ is the

same for type Y. Both a and b can be positive or negative considering the replication error. We note that all the parameters in the matrix Ω can depend on time t in principle. We however treat them as if time independent in this section. A slow time dependence will be introduced in the eigenvalues of matrix Ω in the actual analysis of the data as a kind of adiabatic approximation.

The number of the infected I at time t of our target population will change according to the following formula:

$$\frac{dI}{dt} = \omega_x x - \omega_y y \qquad [3.4]$$

where ω_x and ω_y are constants. The eigenvalues of matrix Ω are given by,

$$det\begin{pmatrix} a(1-Q)-\xi & bR \\ aQ & b(1-R)-\xi \end{pmatrix} = 0 \qquad [3.5]$$

which gives

$$\xi = \frac{1}{2}\left(a(1-Q)+b(1-R) \pm \sqrt{\{a(1-Q)-b(1-R)\}^2+4abQR}\right). \qquad [3.6]$$

In general, ξ is a complex number:

$$\xi = \xi_R \pm i\xi_I \qquad [3.7]$$

with

$$\xi_R = \frac{1}{2}\{a(1-Q)+b(1-R)\} \qquad [3.8a]$$

and

$$\xi_I = \frac{1}{2}\left[\sqrt{-4abQR-\{a(1-Q)-b(1-R)\}^2}\right]. \qquad [3.8b]$$

We get

$$\Omega = P^{-1}\begin{pmatrix} \xi_R + i\xi_I & 0 \\ 0 & \xi_R - i\xi_I \end{pmatrix} P \qquad [3.9]$$

where P is the rows of the eigenvectors of Ω. Eq. [3.3] gives

$$\begin{pmatrix} x \\ y \end{pmatrix} = P^{-1} \begin{pmatrix} A\exp\{\int(\xi_R + i\xi_I)dt\} \\ B\exp\{\int(\xi_R - i\xi_I)dt\} \end{pmatrix}. \qquad [3.10].$$

The general real solution is given as

$$x = C\exp\left(\int_0^t \xi_R dt\right)\sin\left(\int_0^t \xi_I dt + k_1\right) \qquad [3.11a]$$

and

$$y = D\exp\left(\int_0^t \xi_R dt\right)\sin\left(\int_0^t \xi_I dt + k_2\right) \qquad [3.11b]$$

where C and D are constants. We may assume that the time dependence of $\xi_R + i\xi_I$ is not large and we ignore the contributions after the linear term in the power series expansion of ξ in terms of t for the later analysis of the data. Although a more precise treatment such as the perturbation theory developed in particle physics is available, it will be explored elsewhere. We note that we have

$$\tan(k_2 - k_1) = \frac{\xi_I}{\xi_R - a(1-Q)} = \frac{\sqrt{-4abQR - \{a(1-Q) - b(1-R)\}^2}}{-a(1-Q) + b(1-R)} \qquad [3.12a]$$

and

$$D = \left|\frac{bQ}{aR}\right| C. \qquad [3.12b]$$

Principle of error catastrophe We see from Eqs. [3.11] that both x and y have an exponential factor $\exp(\int_0^t \xi_R dt)$. This means (due to Eq. [3.4]) that, for positive ξ_R, the infection number behaves wildly, oscillation becoming larger and larger in its size. For vanishing ξ_R, we have a periodical oscillation just like in regular influenza. The error catastrophe is defined to be the case when we have $\xi_R < 0$. A condition of this to happen for time-independent ξ_R is given by

$$\xi_R = \frac{1}{2}\{a(1-Q) + b(1-R)\} < 0. \qquad [3.13]$$

We assume that Eq. [3.13] is valid for SARS-CoV-2 virus at least for $x > x_0$ (error threshold). There will be periods where this condition is not valid especially at the beginning of a surging of certain variant virus. The change of the infection number given in Eq. [3.4] is written as

$$\frac{dI}{dt} = \omega_x x - \omega_y y = \mathrm{e}^{\xi_R t}[\rho \sin(\xi_I t + k_1) - \lambda \sin(\xi_I t + k_2)] \qquad [3.14a]$$

and

$$\rho = \omega_x C,\ \lambda = \omega_y D. \qquad [3.14b]$$

Here, for simplicity we assumed the eigenvalues to be time independent. Just to see the behavior of dI/dt for a positive or a negative ξ_R, we plot Eq. [3.14] in Fig. 1a and 1b for two cases:

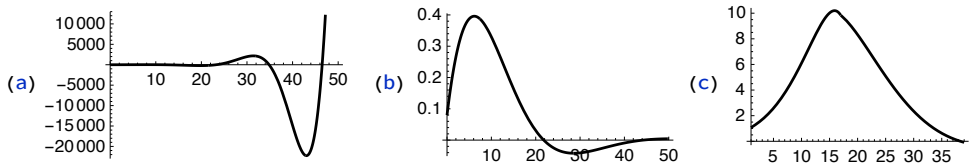

Figure 1. (a) Behavior of Eq. [3.14] for $\xi_R = 0.2, \xi_I = 0.1,\ \rho = 5,\ \lambda = -0.1,$ $k_1 = 0,\ k_2 = 0.83$. (b) Behavior of Eq. [3.14] for $\xi_R = -0.1, \xi_I = 0.1,\ \rho = 1,$ $\lambda = 0.1,\ k_1 = 1.0,\ k_2 = 0$. (c) Eq. [3.14] with $\xi_R = 0.2$ for $0 \le t \le 17$ and $\xi_R = 0.2$ for $17 \le t \le 39$. The other parameters are common: $\xi_I = 0.1, \rho = 1,\ \lambda = 0.1,\ k_1 = 1.0,\ k_2 = 0$.

These values correspond to the following basic parameters:

$$Q = R = \frac{3}{4},\ b = -a + \varepsilon \qquad [3.15a]$$

with

$$\xi_R = \varepsilon(1 - Q) = \frac{\varepsilon}{4} \qquad [3.15b]$$

$$\xi_I = \sqrt{\frac{1}{2}\left\{(a - \frac{\varepsilon}{2})^2 - \frac{3\varepsilon^2}{8}\right\}} \qquad [3.15c]$$

$$\tan(k_2 - k_1) = \frac{\xi_I}{\xi_R - \frac{a}{4}} \qquad [3.15d]$$

and $\varepsilon = 0.8$ for the case (a) and $\varepsilon = -0.4$ for the case (b). We also plot Eq. [3.14a] when ξ_R changes from 0.2 to -0.1 in the middle of peaking of the spreading. Concretely, we choose (a) initially, and (b) right after the peaking of the spreading. We see from these plots that:

(1) There is a maximum value for the change of infection number for either case of $\xi_R > 0$ or $\xi_R < 0$.

(2) Whether the positive ξ_R is kept (a) or it changes to a negative value (b), that is, whether dI/dt chooses the case depicted in Fig. 1c depends on the actual data.

Incidentally, we observe that any country or city has experienced a plural number of peaks in the COVID-19 daily infection number maybe except for China. It shows that each peak is caused by a new variant which is sometimes drastically new, such as δ type or omicron type, and sometimes just a small number of mutations different from the preceding variant.

Incorporating the standard model of epidemiology Our method to combine the above formalism on mutation and the standard model of epidemiology is straight forward as shown here. The standard model (SIR) (Kermack and McKendrick, 1927) is defined to be the following set of differential equations for S (susceptible), I (infected), and R (recovered):

$$dS/dt = -\beta\, IS/N \qquad [3.16a]$$
$$dI/dt = \beta\, IS/N - \gamma I \qquad [3.16b]$$
$$dR/dt = \gamma I\ . \qquad [3.16c]$$

Here $\beta\, IS/N$ is the rate of decrease of the susceptible S, γ is the recovery rate (not to be confused with the symbol of a CoV-2 variant strain) and $S + I + R = N$.

We assume N is a constant (total population of a body we want to discuss). On the other hand, we have Eq. [3.4] which defines the infection numbers in terms of two types of variants x and y defined by Eqs. [3.1] and [3.2]. By regarding Eqs. [3.16a] and [3.16c] as defining S and R, the only useful equation is Eq. [3.16b]:

$$\frac{dI}{dt} = \beta \frac{IS}{N} - \gamma I. \qquad [3.17]$$

Assuming that the most of the members of the group are susceptible, that is, $S/N \approx 1$, we get

$$\frac{dI}{dt} = (\beta - \gamma) I. \qquad [3.18]$$

Our idea is that we additively combine epidemiological Eq. [3.18] with genetical Eq. [3.4]. Unlike Bedford et al. (2021), an implicit assumption is that the infection rate is solely determined by pathogen variants and independent of both the host immune history and possible time-dependent antigenic mutation rates. Then we have the equation

$$\frac{dI}{dt} = \Gamma I + \omega_x x - \omega_y y \qquad [3.19]$$

with

$$\Gamma = \beta - \gamma. \qquad [3.20]$$

The daily infection number is defined as

$$\text{Daily infection number } \equiv \frac{dI}{dt}. \qquad [3.21]$$

We solve linear differential Eq. [3.19] for I using the ordinary method of variation of parameters (in this case it is $f(t)$), or we assume I in the form of $I(t) = f(t)\Gamma e^{\Gamma t}$, substitute it into both sides, and obtain the result of $f(t)$:

$$\frac{dI}{dt} = \frac{df(t)}{dt}\Gamma\, e^{\Gamma t} + \Gamma I = \Gamma I + \omega_x x - \omega_y$$

$$= \Gamma e^{\Gamma t}\left[f(0) + \int_0^t (\omega_x x - \omega_y) e^{-\Gamma\tau} d\tau\right] + \omega_x x - \omega_y y$$

$$= \Gamma e^{\Gamma t}\left[f(0) + \rho\left\{H\left(t + \frac{k_1}{\xi_I}\right) - H\left(\frac{k_1}{\xi_I}\right)\right\} - \lambda\left\{H\left(t + \frac{k_2}{\xi_I}\right) - H\left(\frac{k_2}{\xi_I}\right)\right\}\right] + \omega_x x - \omega_y y \qquad [3.22]$$

where we used Eq. [3.14] and defined $H(x)$ as

$$H(x) = \frac{e^{(\xi_R - \Gamma)x}(-\xi_I \cos[\xi_I x] + (\xi_R - \Gamma)\sin[\xi_I x])}{(\xi_R - \Gamma)^2 + {\xi_I}^2} \ . \qquad [3.23]$$

4. ANALYSIS OF ACTUAL DATA

In this section, we analyze some data based on Eq. [3.22] with [3.23]. We see that the formula has the generic form:

$$\frac{dI}{dt} = c_1 e^{\Gamma t} + e^{\xi_R t}\{c_2 \sin(\xi_I t) + c_3 \cos(\xi_I t)\} \qquad [4.1]$$

where c_i $(i = 1, 2, 3)$ are certain constants. We note that ξ_R and ξ_I are linear functions of t as dicussed in the last section. We see that the first term on the right-hand side represents the contribution of the non-genetic factor such as vaccination, body distancing and PCR testing. The second term represents the genetic effect.

Tokyo 5th wave We analyze the Tokyo 5th wave data just to see how our new way of analysis works especially in comparison with the standard epidemiology model in Eqs. [3.16]. The Tokyo data on daily infection numbers in the 5th wave period are given in https://www.google.com/search?q=tokyo+data+on+covid-19&oq=tokyo+data&aqs=chrome.1.69i57j35i39j0i512l3j0i30l5.9499j0j15&sourceid = chrome&ie = UTF-8. From the data supplement 1, we have:

Data list for Tokyo 5th wave ={453, 619, 317, 716, 896, 830, 1008, 1359, 3177,

2195, 4498, 4989, 4377, 4392, 4227, 3168, 968, 1273, 763, 253, 299, 200}.

[4.2]

We fit the data with Eq. [4.1] using the nonlinear fitting of Mathematica (Fig. 2). For simplicity, we subsequently denote ξ_R by R and ξ_I by J. It turned out that we cannot fit the data of the entire period of the 5th wave if we keep the time independent R. This is reasonable because the error threshold ($R = 0$) cannot be reached at the start of the spreading. It will be achieved sometime during the spreading process.

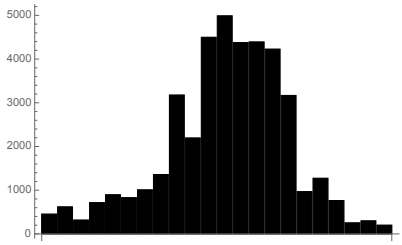

Figure 2. Tokyo daily infection numbers from June 18 to October 1, 2021, (against the horizontal axis x measured in units of 5 days, or $x = t/5$).

We use two strategies to fit the data:
(1) We split the data into three periods and fit the data in each period separately.
(2) We assume a simple form of time dependence for R. The advantage of this method is that we can calculate from the data when the error threshold is reached.

We start with the method (1): We divide the period of the Tokyo 5th wave into three periods: (a) first 40 days, (b) middle 20 days, and (c) later 40 days. The result of the fitting is as follows:

(a) period:

$$\frac{dI}{dt} = 0.00018e^{0.406t} - 133.7e^{0.039t}\sin[0.016t] + 419.2e^{0.039t}\cos[0.016t]. \quad [4.3]$$

The curve [4.3] is depicted in Fig. 3a. Note that R (0.039) is positive indicating that the error threshold has not been reached yet.

(b) period: We get

$$\frac{dI}{dt} = 34.49e^{-0.044t} - 23.68e^{0.0998t}\sin[0.0196t] + 87.13e^{0.0998t}\cos[0.0196t]. \quad [4.4]$$

Note that R (0.0998) is even larger than the initial period value of 0.039, and it still keeps below the threshold (Fig. 3b).

(c) period: In this case we get

$$\frac{dI}{dt} = 3550e^{-0.0026t} + 3171e^{-0.0093t}\sin[0.056t] - 5652e^{-0.0093t}\cos[0.056t]. \quad [4.5]$$

We see that R (-0.0093) is negative in this period satsfying the error catastrphe condition. Eq. [4.5] is depicted in Fig. 3c.

We learned above that the value of R is dependent on which period we consider. This suggests that we may assume some simple form of time dependence for R and also for Γ in Eq. [4.2] (see the discussion following Eqs. [3.3] and [3.11a, b] as well as the Appendix concerning symbiosis). The form we assume is the following:

$$R = R_0(1 - At) \quad [4.6]$$

$$\Gamma = \Gamma_0(1 - Bt). \quad [4.7]$$

Then, we fit the entire data with a single function:

$$c_1 e^{\Gamma_0(1-Bt)t} + e^{R_0(1-At)t}\{c_2\sin[Jt] + c_3\cos[Jt]\}. \quad [4.8]$$

The result of the fitting is given as:

$$\frac{dI}{dt} = 453e^{0.024(1-0.0099t)t} + 3.77e^{0.216(1-0.0078t)t}\sin[0.041t] \\ -2.32e^{0.216(1-0.0078t)t}\cos[0.041t]. \quad [4.9]$$

We see from Eq. [4.9] that R changes from positive to negative during the 5th wave period. We can calculate the time of going over the error threshold from:

$$R = 0.216(1 - 0.0078t_t) = 0 \qquad [4.10]$$

namely, $t_t = 128$ days after the initiation of the δ variant spreading. Although this value of t_t roughly indicates the starting of the decay of the spreading, we see from Eq. [4.9] that other factors and even the non-genetic first term contribute to the precise determination of the maximum position of the infection numbers. Fitting of the entire period by Eq. [4.9] is shown in Fig. 3d. The fitting describes clearly the battle ground of virus vs human immune system. Initially the virus is winning but eventually it reaches the maximum spreading power. The APOBEC3 deaminases involved in the human innate immune system (Robson et al., 2020; Simmonds, 2020; Ratcliff and Simmonds, 2021) keep attacking the virus RNA to keep mutating, which cannot avoid more and more errors so that the spreading power starts to shrink.

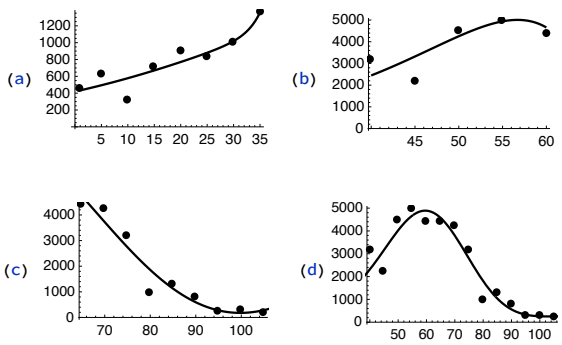

Figure 3. (a) Daily infection number of Tokyo at the initial 40 days of the 5th wave expressed in Eq. [4.3]. (b) Tokyo infection number in the middle 20 days expressed in Eq. [4.4]. (c) Later stage of the Tokyo 5th wave as expressed in Eq. [4.5]. (d) Tokyo 5th wave fit by Eq. [4.9].

5. ANALYSIS OF DATA FROM SOUTH AFRICA AND JAPAN

We learned in section 4 that it is possible to fit the data of an entire period of a certain wave if we take into account the time dependence of R and Γ (infection rate β minus recovery rate γ in Eq. [3.20]). We adopt this strategy rather than dividing a period into several sub-periods.

The omicron data are still not sufficient and, therefore, the following analysis should be taken as preliminary rather than final. The omicron variant was first found in South Africa. Therefore, we begin our analysis of the omicron variant in South Africa. The data [https://www.worldometers.info/coronavirus/country/south-africa/] from November 29, 2021 to January 13, 2022 is given as follows (Data supplement 2):

Data list for South Africa={2273, 4393, 8561, 15535, 16055, 16366, 10000, 6381, 13147, 19842, 22383, 19017, 17153, 37875, 13288, 20000, 26389, 24785, 20713, 16080, 15465, 8511, 15423, 21098, 21156, 13000, 14824, 5603, 3778, 6000, 9020, 12978, 11754, 9793, 6000, 3076, 8078, 11106, 9858, 9259, 7759, 4482, 2409, 4500, 6760, 5917}. [5.1]

The data is depicted in Fig. 4a. We fit the data with a function that is exactly the same as the one in Eq. [4.8], which is used to explain the entire period of the Tokyo 5th wave. The result of the fitting gives:

$$\frac{dI}{dt} = 78.71e^{0.38(1-0.0156t)t} + 207.4e^{0.206(1-0.021t)t}\sin[0.072t] + 4548e^{0.201(1-0.021t)t}\cos[0.072t]. \quad [5.2]$$

We see from Eq. [5.2] that

$$R = 0.201\,(1 - 0.021t). \quad [5.3]$$

Eq. [5.3] satisfies the error catastrophe condition of $R < 0$ after $t_t = 48$ days. This value of 48 days is much smaller than the Tokyo 5th wave (supposed to be the δ type) value of $t_t = 128$ days, indicating that omicron infection is more powerful and in turn it reaches the error threshold much faster than the δ type. We show the curve described by Eq. [5.2] in Fig. 4b. We extend Fig. 4b to much later time as in Fig. 4c. We see from Figs. 4b, c that omicron variant infection has a much steeper increase but also sharp decrease due to the earlier arrival at the error threshold.

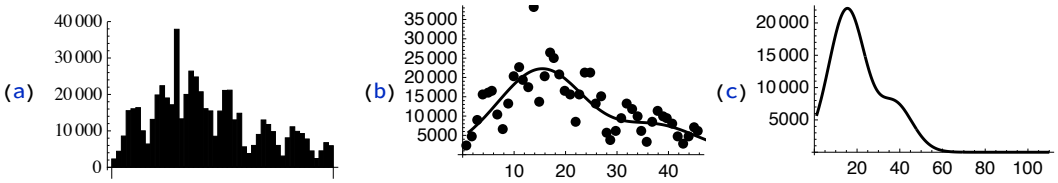

Figure 4. (a) South Africa data of daily infection shown in Eq. [5.1]. (b) South Africa omicron data fit by Eq. [5.2]. (c) Extension of (b) to much later time.

Universality of the parameters ξ_R (= R) and ξ_I (= J) Before fitting the Japan data, we discuss the nature of the parameters ξ_R and ξ_I. As shown above the time to reach the error threshold is calculated using the time dependence of ξ_R. We obtained the value $t_t = 128$ days for the δ type using the Tokyo data and $t_t = 48$ days for the omicron variant using the South Africa data. We assume that these values depend on the variants we consider, but not or very little on any other external effect such as vaccination, social distancing etc.. The reason is that ξ_R and ξ_I are genetic parameters but not the epidemic parameters. They are more or less inherent in the RNA of the virus and can also depend on the genetically determined human immune system. The other parameters in Eq. [4.1] can depend on vaccination, social distancing, PCR testing followed by segregation etc.. The above argument allows us to use the values of ξ_R and ξ_I obtained from the South Africa data to analyze the omicron data of other regions, although there may be some difference of values of ξ_R and ξ_I due to population structure such as how much of particular genomic segments of Neanderthals introgressing into the DNA of the group we deal with (Zeberg and Pääbo, 2020). In the following analysis of the Japan data, we use the South Africa ξ_R and ξ_I for simplicity.

The validity or the usefulness of the universality of ξ_R and ξ_I can be checked by the analysis using such index as the AIC. Along the same line, we can also consider (calculate the AIC of) the limiting case of all the genetic part of the parameters vanishing. These works are left to our future publication.

Japan's case -- 6th wave We analyze the Japan data of daily infection number starting December 30, 2021 to January 13, 2022 when the omicron variant was starting to dominate. We use the same ξ_R and ξ_I obtained in the South Africa analysis. The data for the 6th wave of Japan is given in the data supplement 3 as,

Data for 6th wave of Japan = {{1, 466}, {2, 490}, {3, 463}, {4, 504}, {5, 600}, {6,

783}, {7, 1256}, {8, 2506}, {9, 4194}, {10, 5983}, {11, 7930}, {12, 8144}, {13, 6394}, {14, 9000}, {15, 12243}, {16, 22045}, {17, 25742}}. [5.4]

We fit this with the following function which is exactly same as the one used in the South Africa analysis with an unjustified assumption that even the non-genetic parameter (B in Eq. [4.7]) is universal:

$$\frac{dI}{dt} = c_1 e^{\Gamma_0(1-0.0156\mathrm{t})\mathrm{t}} + c_2 \mathrm{e}^{0.2015(1-0.0207t)t} \sin[0.072t] + c_3 \mathrm{e}^{0.2015(1-0.0207t)t} \cos[0.072t]. \quad [5.5]$$

The result of the fitting gives

$$\frac{dI}{dt} = 0.00072 e^{1.158(1-0.0156t)t} + 2116 \mathrm{e}^{0.2015(1-0.0207t)t} \sin[0.072t] + 496.1 \mathrm{e}^{0.2015(1-0.0207t)t} \cos[0.072t]. \quad [5.6]$$

The data [5.4] and Eq. [5.6] are plotted in Fig. 5a. From Fig. 5b, we see that Japan's daily infection number will go up to approximately 100,000 and then it starts decreasing. Our analysis of the Japan data (dataJPN1 in [5.4]) from December 30, 2021 to January 13, 2022 correctly predicted the maximum daily infection number to be approximately 100,000. But the actual speed of the decline from the peak number is much less than predicted by the analysis. This is due to the unjustified assumption that even the non-genetic parameters are universal.

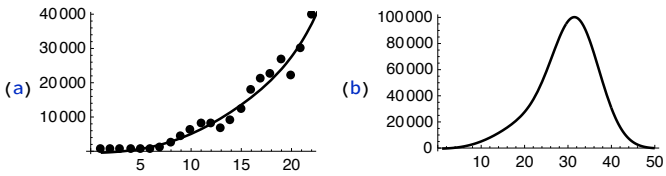

Figure 5. (a) Japanese daily infection number fitted by Eq. [5.6]. (b) Extension of (a) to later time.

Analysis of later data of Japan We investigate the situation by analyzing the data from January to March, 2022. Here we drop the unjustified assumption that the non-genetic parameters are universal. The data from January 1 to March 31 of the

infection number (every 3 days) is given in the data supplement 4 as,

Later data of Japan = {{1, 2505}, {2, 8144}, {3, 6829}, {4, 2124123351}, {5, 44638}, {6, 47176}, {7, 69736}, {8, 82159}, {9, 93388} {10, 94431}, {11, 7170},{12, 10097}, {13, 80234}, {14, 95603}, {15, 77135 }, {16, 66373}, {17, 71254}, {18, 55243}, {19, 70958}, {20, 55417}, {21, 62747}, {22, 52002}, {23, 50781}, {24, 49210}, {25, 29657}, {26, 47937}, {27, 32916}, {28, 53544}}. [5.7]

We fit the data with the following function:

$$\frac{dI}{dt} = c_1 \mathrm{e}^{\Gamma_0(1-B\times 3t)3t} + c_2 \mathrm{e}^{0.2015(1-0.0207\times 3t)\times 3t} \sin[0.072 \times 3t] + c_3 \mathrm{e}^{0.2015(1-0.0207\times 3t)3t} \cos[0.072 \times 3t]. \quad [5.8]$$

The meaning of the right-hand side in Eq. [5.8] is the following:
(1) Compared to the previous fitting curve given in Eq. [5.5], we see that we use "$3t$" rather than "t" to designate the time variable (here we have data points of every 3 days).
(2) Otherwise, the only difference is the insertion of the parameter B in the first term (social distancing contribution). This is the most important deviation from Eq. [5.5] where the unjustified assumption of the non-genetic term was made.
(3) The universality of the mutation contribution is maintained: The second and third terms are the same as in Eq. [5.5].

The result of the fitting gives:

$$\frac{dI}{dt} = 6882\mathrm{e}^{0.2431(1-0.0269t)t} + 2886\mathrm{e}^{0.6046(1-0.0621t)t} \sin[0.216t] - 3106\mathrm{e}^{0.6046(1-0.0621t)t} \cos[0.216t]. \quad [5.9]$$

This non-linear fitting is shown in Fig. 6a. We extend the function in Eq. [4.9] to $t = 50$ (150 days after January 1, 2022, the end of May). It is depicted in Fig. 6b.

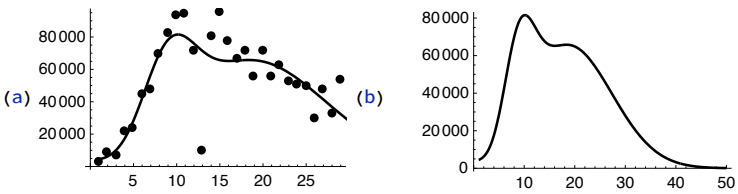

Figure 6. (a) Fitting of the Japan data of daily infection from January 1 to April 2 fitted by Eq. [5.9]. (b) Daily infection number of Japan from early January to the end of May given in Eq. [5.9].

Tentative summary of the analysis of the later data of Japan (1) As one can see by comparing Eqs. [5.6] and [5.9], the difference between the previous analysis and the present one is mostly in the first term representing the contribution of the social distancing effect. The previous best fitting ended up in minimizing the first term, whereas the current fitting has much larger contribution from this term.
(2) The exponential factor of the first term in the current analysis $[\exp\{\frac{0.6046}{3}\left(1-\frac{0.0621}{3}\times 3t\right)3t\}]$ behaves more or less the same way as that of the previous analysis $[\exp\{0.2015(1-0.0207t)t\}]$ indicating the structure of the social distancing effect was correctly taken into account in both analyses. The difference is in the overall factor of the first term (0.00072 vs 6882).
(3) As one can see from Fig. 6b, there will be no 7th wave in Japan if our analyses are correct. Unfortunately, however, it appears that before omicron BA. 2 completely disappears, it was gradually replaced by a new omicron variant strain (BA. 5), which has caused the 7th wave in 2022 summer Japan.

One of the authors (H. S.) would like to express his sincere gratitude to Professor Toshimichi Ikemura for his variable discussions and guidance.

REFERENCES

Bedford, T., Rambaut, A., and Pascual, M. (2012) Canalization of the evolutionary trajectory of the human influenza virus. BMC biology 10: 38.

Eigen, M., (1971) Self-organization of matter and the evolution of biological macro-molecules. Naturwissenschaften 58: 465-523.

Kermack, W. O., and McKendrick, A. G. (1927) A contribution to the mathematical theory of epidemics. Proc. R. Soc. Lond. A 115: 700-721.

Kucharski, A. J., Andreasen, V., and Gog, J. R. (2016) Capturing the dynamics of pathogens with many strains. J. Math. Biol. 72: 1-24.

Miller, J. K., Elenberg, K., and Dubrawski, A. (2022) Forecasting emergence of COVID-19 variants of concern. PLoS ONE 17(2): e0264198.

Ratliff, J., and Simmonds, P. (2021) Potential APOBEC-mediated RNA editing of the genomes of SARS-CoV-2 and other coronaviruses and its impact on their longer term evolution. Virology 556: 62-72.

Robson F., Khan, K. S., Paris, C., Demirbag, S., Barfuss, P., Rocchi, P., and Ng, W. L. (2020) Coronavirus RNA proofreading: Molecular basis and therapeutic targeting. Mol. Cell 80: 1136-1138.

Rohani, P., Green, C. J., Mantilla-Beniers, N. B., and Grenfell, B. T. (2002) Ecological interference between fatal diseases. Nature 422: 885-888.

Sadeghpour, S., Khodaee, S., Rahnama, M., Rahimi, H., and Ebrahimi, D. (2021) Human APOBEC3 variations and viral Infection. Viruses 13: 1366. https://doi.org/10.3390/v13071366.

Simmonds, P. (2020) Rampant C→U hypermutation in the genomes of SARS-CoV-2 and other coronaviruses: Causes and consequences for their short- and long-term evolutionary trajectories. mSphere 5: e00408-20.

Smith, E. C., Blanc, H., Vignuzzi, M., and Denison, M. R. (2013) Coronaviruses lacking exoribonuclease activity are susceptible to lethal mutagenesis: Evidence for proofreading and potential therapeutics. PLoS Pathog 9(8): e1003565. doi: 10.1371/journal.ppat.1003565.

Sugawara, H. (2021a) On the effectiveness of the search and find method to suppress spread of SARS-CoV-2. Proc. Jpn. Acad. Ser. B 97(1): 22-49.

Sugawara, H. (2021b) Effect of vaccination against COVID-19 spreading. Proc. Jpn. Acad. Ser. 97(9): 543-558.

Thakur, V., Bhola, S., Thakur, P., Patel, S. K. S., Kulshrestha, S., Ratho, R. K., and Kumar, P. (2022) Waves and variants of SARS-CoV-2: understanding

the causes and effect of the COVID-19 catastrophe. Infection 50: 309-325.
Zeberg, H., and S. Pääbo (2020) The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature 587: 610-612.

APPENDIX: SYMBIOSIS OR TOTAL EXTINCTION

Our definition of error catastrophe discussed in section 2 is given in terms of the real part of the eigenvalue of the matrix which determines the numbers of variants which contribute positively or negatively to the infection of a given virus. The eigenvalue is time dependent in general. If the real part is positive, the positive variant type dominates and vice versa. The error catastrophe occurs when the real part becomes zero at certain stage. If the negativity persists, we will reach the total extinction of the virus.

Now suppose that the real part of the eigenvalue stays zero after reaching the error catastrophe. This is the situation of the symbiosis due to our definition. Then our basic Eq. [4.1] becomes,

$$\frac{dI}{dt} = c_1 e^{\Gamma t} + c_2 \sin(\xi_I t) + c_3 \cos(\xi_I t). \qquad [A.1]$$

The first term is the contribution from the artificial measure such as vaccination, social distancing etc.. The rest shows the mutation effect. The value of Γ can be negative without any of the artificial measures corresponding to the weakened infection power of the virus. It is possible that we still need vaccination etc. to make it negative.

On the mutation effect, we see that the daily infection number oscillates as a function of time if the imaginary part of $\xi_I\ (= J)$ is non-zero. This implies that, in case of symbiosis, there will be maximum and minimum in the infection rate somewhat similar to the situation where the infection rate reached the error catastrophe. But the difference is that the former is realized due to the non-zero property of ξ_I, whereas the latter is due to the vanishing of $\xi_R\ (= R)$.

This means that there could be ups and downs of daily infection numbers in the case of symbiosis, as in the case of seasonal influenza, and this phenomenon is not the same as in the case of error catastrophe.

Data Supplement

1. Data list for Tokyo 5th wave

Our format of the data is such that we list the number of patients starting from June 18 to October 1, 2021 every 5 days:

Data list for Tokyo 5th wave ={453, 619, 317, 716, 896, 830, 1008, 1359, 3177, 2195, 4498, 4989, 4377, 4392, 4227, 3168, 968, 1273, 763, 253, 299, 200}.

2. Data list for South Africa

Our format of the data here is such that we list the number of patients starting from November 29, 2021 to January 13, 2022 every day:

Data list for South Africa={2273, 4393, 8561, 15535, 16055, 16366, 10000, 6381, 13147, 19842, 22383, 19017, 17153, 37875, 13288, 20000, 26389, 24785, 20713, 16080, 15465, 8511, 15423, 21098, 21156, 13000, 14824, 5603, 3778, 6000, 9020, 12978, 11754, 9793, 6000, 3076, 8078, 11106, 9858, 9259, 7759, 4482, 2409, 4500, 6760, 5917}.

3. Data for 6th wave of Japan

Here the same data format as in the case of South Africa is used except that we add the numbering of dates on the first column of the list:

Data for 6th wave of Japan = {{1, 466}, {2, 490}, {3, 463}, {4, 504}, {5, 600}, {6, 783}, {7, 1256}, {8, 2506}, {9, 4194}, {10, 5983}, {11, 7930}, {12, 8144}, {13, 6394}, {14, 9000}, {15, 12243}, {16, 22045}, {17, 25742}}.

4. The Japan data from January 1 to March 31 of the infection number (every 3 days) (https://www.worldometers.info/coronavirus//country/japan/) (using the same format as in the case of 6th wave):

Later data of Japan = {{1, 2505}, {2, 8144}, {3, 6829}, {4, 2124123351}, {5, 44638}, {6, 47176}, {7, 69736}, {8, 82159}, {9, 93388} {10, 94431}, {11, 7170},{12, 10097}, {13, 80234}, {14, 95603}, {15, 77135 }, {16, 66373}, {17, 71254}, {18, 55243}, {19, 70958}, {20, 55417}, {21, 62747}, {22, 52002}, {23, 50781}, {24, 49210}, {25, 29657}, {26, 47937}, {27, 32916}, {28, 53544}}.