The rich phase structure of a mutator model

David B. Saakian[1,2],* Tatiana Yakushkina[3],† and Chin-Kun Hu[1‡]
[1] *Institute of Physics, Academia Sinica, Nankang, Taipei 11529, Taiwan*
[2] *A.I. Alikhanyan National Science Laboratory (Yerevan Physics Institute) Foundation, 2 Alikhanian Brothers St., Yerevan 375036, Armenia and*
[3] *National Research University Higher School of Economics, Moscow, 101000, Myasnitskaya Str. 20, Russia,*
(Dated: September 25, 2015)

We investigate a biological evolution model where the genome contains a specific gene which can be in a mutator allele increasing mutation rate or changing the fitness landscape. We calculate the phase structure (important for cancer), the mean fitness, the surplus and what allele takes over the population (important for virology). There are mutator phase, mixed phase and, depending on fitness landscape, the non-selective fitness phase. In the mixed phase, either normal allele or mutator dominates, depending on genome length. We solve exactly the model for linear Malthusian fitness function. We conclude that the random fitness landscape can be a proper choice for the observed mutator phenomenon. Our theoretical findings have a direct biomedical impact. We find 4 different situations from the bio-medical perspectives. One should clearly distinguish the increase of mutation rates in two parts of genome: only some combination of these increases can push the system to non-selective phase, potentially related to eradiation of tumors.

The concepts of genome instability [1–6] and clonal evolutionary dynamics [7–9] are among the key ideas in understanding cause and behavior of cancer. The first phenomenon has been considered as one of hallmarks of cancer [6], moreover it plays an important role for bacteria evolution as well [10–14]. The genome instability is one of the focuses of physics community working in oncology [15] due to its serious biomedical impact. Clonal (asexual) evolution models are covered in [17–20], while for cancer analysis it necessary to use more complex nonlinear models [7–9, 16]. Models with relatively simple evolutionary dynamics can be applied to cancers with one or two mutations, e.g. inherited retinoblastoma, however the most cancers require three or more driver mutations. For the latter case a mutator phenotype is inevitable. It has been proposed that there is a special mechanism of genome instability to create a large number of mutations. The mutator phenomenon can be realized by the mutation of the gene, responsible for a genome stability, from the normal allele into abnormal allele and as a result there is a substantial increase in the mutation rate of the genome. There have been some attempts to construct and solve the evolutionary dynamics with a mutator gene [11, 12]. In [12] has been found the solution of the mutator model in case of simplest (linear fitness function in Crow-Kimura model version). In [21] has been solved the infinite population model, similar to the Crow-Kimura model, only with constant mutation rates (regardless the Hamming classes), see also the recent review [22]. Especially important is the new evolutionary dynamics phase found first in [12]. The mutator model is useful to describe the modern approach to the cancer as a process similar to RNA virus evolution [4] and the recent theory of cancer as a reverted evolution [23]. Recently a key aspect of asexual evolution, an error threshold, has been enlarged for the cancer case [5, 16], and the lethal mutagenisis has been suggested to cure the cancer. For such biomedical applications it is important to calculate the error threshold. In case of bacteria evolution it is important to calculate the probability of fixation for an abnormal mutator allele. There have been several theoretical works to consider evolutionary dynamics in case of mutator. In [11] a microscopic model (a mixture of Crow-Kimura and Eigen model) with a linear fitness function has been considered and some approximate estimates have been done for the steady state distribution and the dynamics of the mutator fraction in population. In [12] has been considered the discrete time Eigen model. There was a very serious finding in [12]. In case of small transition rates (normal allele to mutator allele), the system is in the "mixed phase", where there is a small fraction of abnormal mutator allele, and at higher transition rates there is a transition into mutator phase where the vast majority of population has a mutator allele.

In this work we give the first exact (at infinite genome length limit case) and comprehensive solution of the model of evolutionary dynamics with mutator. We find the exact solution for the mean fitness for the rather general fitness function, even in multidimensional case, clarify and calculate the order parameters, the phases and sub-phases of the model. Our quantitative results about the borders of different phases in Fig. 2 are important for the biomedicine in

*Electronic address: saakian@yerphi.am
†Electronic address: tyakushkina@hse.ru
‡Electronic address: huck@phys.sinica.edu.tw

view of the suggestion of [4] pushing the tumor from the mixed phase to the non-selective phase to cure the disease. The large (infinite population limit) is the main assumption of our model, then we solve the model in two cases, very popular in population genetics: assuming either a symmetry of fitness landscape (the fitness landscape depends on the total number of mutations from the reference or references sequences) or the randomness of fitnesess. In this work we will construct the mutator gene model on the basis of Crow-Kimura (parallel) mutation-selection model [13, 17], later consider the solution of the Eigen model version of the mutator [12]. The Crow-Kimura model, used in our work, is equivalent in the large genome limit to discrete time Eigen model of [12], see [24], and is very close to branching processes. This parallel selection-mutation model [17] is widely investigated last decades as a microscopic model for the virus evolution, altogether with the Eigen's model [18, 20]. These models are described via a system of non-linear master equations, which can be mapped to the chain of linear ODE plus some nonlinear transformation. The mean fitness of the Crow-Kimura model has been calculated using the algebraic methods [25], as well as the Hamilton-Jacobi equation (HJE) method [26, 27]. Later the exact dynamics has been derived using the HJE [28].

In the Crow-Kimura model with a symmetric landscape, the genome is described as a chain of N letters (genes), taking values ± 1. There is a mutation rate μ per letter. We take the sequence with all "+" letters as a reference sequence. All the genomes with the same total number l of "–" letters (l point mutations from the reference sequence) have the same fitnesses $r_l \equiv f(1-2l/N)$, where we define the function $f(x)$ as a fitness function, with $x = 1-2l/N$. For the symmetric initial distribution one can describe the model via differential equations of $N+1$ variables P_l, corresponding to the total probabilities of the viruses with the l mutations in the genome (Hamming classes). Here we will use the Crow-Kimura model to construct the evolution model in case of mutator.

Results

The model with symmetric fitness landscape.

In this paper we consider a modification of the Crow-Kimura model for evolutionary process under the influence of a mutator gene. The chain of $(N+1)$ genes defines the genome, which can be formally subdivided into two parts: a regular part with N genes and a mutator gene, where any mutation changes the fitness and the mutation rate in the regular part of genome. We represent each gene with 2 alleles by $s_\tau = \pm 1, \tau = 0, \ldots, N$. The regular set of genes is denoted by the sequence $S_i = (s_1, \ldots, s_N), \quad i = 1, \ldots, 2^N$. The mutator gene in the state $s_0 = +1$ determines normal dynamics of the regular part with mutation rate μ_1 (wild type), but when the state changes to $s_0 = -1$ evolutionary process switches to different regime with mutation rate μ_2 (mutant). The mutation rate for the mutator gene itself is denoted by α_1 ($s_0 : +1 \to -1$) and α_2 ($s_0 : -1 \to +1$). We assume that fitness landscape is symmetric, which

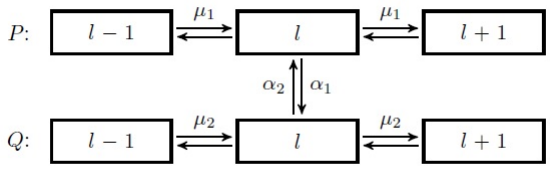

FIG. 1: The scheme of available transitions for the system states (arrows denote transitions). Upper chain corresponds to the genome without a mutator allele; the lower chain corresponds to the genome with a mutator allele. l is a number of mutations in the regular part of genome.

means it depends solely on the number of mutations l from the reference sequence (without loss of generality we take reference sequence to be $S = (+1, \ldots, +1)$). For this purpose it is natural to use the notion of Hamming distance between S and $S_i : l \equiv d_{1i} = (N - \Sigma_{\tau=1}^{N} s_\tau)/2$, since all sequences in the same Hamming class have the same value of fitness. To facilitate further analysis, we introduce the variable for the mean value $x_l = \sum_{\tau=1}^{N} s_\tau / N, \quad x_l \in [-1, 1]$. The state of the system is represented by two probability distributions. We use $P_l(t)$ to denote relative frequency of normal genome sequences with l mutations (in the l–th Hamming class) at the time moment t and $Q_l(t)$ — for relative frequency of sequences with increased mutation rate with l mutations. All admissible transitions between system states could be seen in Fig.1 as arrows.

To describe the evolution of probability distribution under consideration, we use the following system of $2(N+1)$

differential equations:

$$\frac{dP_l(t)}{Ndt} = \alpha_2 Q_l + P_l\left(f(x_l) - (\mu_1+\alpha_1)\right) + \mu_1\left(P_{l-1}\frac{N-l+1}{N} + P_{l+1}\frac{l+1}{N}\right) - P_l R,$$
$$\frac{dQ_l(t)}{Ndt} = \alpha_1 P_l + Q_l\left(g(x_l) - (\mu_2+\alpha_2)\right) + \mu_2\left(Q_{l-1}\frac{N-l+1}{N} + Q_{l+1}\frac{l+1}{N}\right) - Q_l R,$$
$$R(t) = \sum_l (P_l(t) f(x_l) + Q_l(t) g(x_l)),$$
$$x_l = 1 - 2l/N, \quad 0 \le l \le N. \tag{1}$$

Here $f(x_l)$ is a fitness function for the regular part with normal mutator gene and $g(x_l)$ – for the sequence with abnormal mutator allele. As long as changes in mutator gene affect significally both fitness and mutation rate in the regular part, fitness functions $f(x_l)$ and $g(x_l)$ can be different and μ_2 is 10–100 times larger than μ_1. One can note that the coefficients $(N-l+1)$ and $(l+1)$ appear in Eq.(1) transition terms according to combinatorial formulas for Hamming class probabilities [25, 29]. To investigate the system (1) we drop the nonlinear terms proportional to R and apply a nonlinear transformation: $P_l \to \frac{P_l}{\sum_l (P_l+Q_l)}, Q_l \to \frac{Q_l}{\sum_l (P_l+Q_l)}$. This substitution gives a system which is equivalent to the nonlinear differential equation [30].

In this paper we focus on the following characteristics of the model at the steady state: the mean fitness R, the surplus s (the expected value of s_τ), the surplus for the fraction of population with normal mutator allele s_1, and for the abnormal mutator allele s_2, the probability of the sub-population q with increased mutation rate:

$$R = \frac{\sum_l (P_l f(x_l) + Q_l g(x_l))}{\sum_l (P_l + Q_l)}; s = \frac{\sum_l (P_l + Q_l) x_l}{\sum_l (P_l + Q_l)}$$
$$s_1 = \frac{\sum_l P_l x_l}{\sum_l P_l}; s_2 = \frac{\sum_l Q_l x_l}{\sum_l Q_l}; q = \frac{\sum_l Q_l}{\sum_l (P_l + Q_l)}. \tag{2}$$

The maximum of the distributions P_l and Q_l are attained at points s_1 and s_2, respectively [28]. Therefore the value of the mean fitness for wild type sequences is equal to $f(s_1)$ and for mutant type — to $g(s_2)$. Assuming a smooth distribution, we obtain from (2):

$$R = q g(s_2) + (1-q) f(s_1). \tag{3}$$

To calculate the mean fitness in growing populations is crucial for understanding the evolutionary dynamics and is a primary concern of this investigation. Depending on the values of main parameters of the model, we can obtain different analytical expressions for the mean fitness, which correspond to different phases of the system. The key problem is to solve first the infinite N limit, and clarify the following items:

- 1. Does $s > 0$?
- 2. Does $s_1 = s_2$?
- 3. Does $q < 1$?
- 4. Are there large finite N corrections for the q or R at $N \sim 10000$?

While the item 1 (the existence of the selective phase) is well known [18, 20], the other criteria are new.

We found that there are possible situations with both negative and positive answers to these four criteria. The most serious differentiation of the phases is according to second criteria. We first investigate the case $s_1 = s_2$, later the more advanced situation with a solution $s_1 \neq s_2$. The fourth criteria is also important. Contrary to the ordinary evolution models from quasi-species, here the evolutionary picture depends on the genome length, and the infinite genome limit is not valid for the calculation of some order parameters as before.

The case of forward and backward transitions of the mutator gene. Let us consider a smooth solution of Eq.(1) at the limit $N \to \infty$ after simplification mentioned in the previous section. We assume the following ansatz [26, 28, 31]:

$$P_l(t) \equiv P(x,t) = v_1(x,t) e^{N u(x,t)}; \quad Q_l(t) \equiv Q(x,t) = v_2(x,t) e^{N u(x,t)}, \tag{4}$$

where we denoted $x = 1 - 2l/N$ for large values of N. This form of expressions for P_l and Q_l allows to apply HJE method to the system of equations (1) (see Methods section for details). We obtain the following formulas for the mean fitness R using the potential function $V_\pm$ [28]:

$$\begin{aligned} R &= \max_x [V_+(x)], \\ V_\pm(x) &= \frac{f(x)+g(x)}{2} - \frac{1+\mu}{2} - \frac{\alpha_1+\alpha_2}{2} + \frac{1+\mu}{2}\sqrt{1-x^2} \pm \frac{1}{2}\sqrt{A(x)^2 + 4\alpha_1\alpha_2}, \end{aligned} \tag{5}$$

here we use $\mu_1 = 1,\ \mu_2 = \mu$, and $A(x) = f(x) - g(x) - \alpha_1 + \alpha_2 + (1-\mu)(\sqrt{1-x^2} - 1)$. The numerical simulation for the system (1) supports well analytical expression (5), as can be seen from the Table I.

Having the expression for the mean fitness R, we can calculate the surplus of distribution $s = s_1 = s_2$ from the following equation (see Methods):

$$R = \frac{f(s)+g(s)-\alpha_1-\alpha_2}{2} + \frac{1}{2}\sqrt{(f(s)-g(s)-\alpha_1+\alpha_2)^2 + 4\alpha_1\alpha_2}. \tag{6}$$

At the time limit $t \to \infty$ we denote $v_k(x) \equiv v_k(x,t), k = 1,2$. Assuming $s_1 = s_2$ and putting the solution of R and s, we get the following system of equations for the surplus s and ratio $\frac{v_2(s)}{v_1(s)}$ in the steady state (see Methods) :

$$\begin{aligned} R &= f(s) - \alpha_1 + \frac{v_2(s)}{v_1(s)}\alpha_2, \\ R &= g(s) - \alpha_2 + \frac{v_1(s)}{v_2(s)}\alpha_1. \end{aligned} \tag{7}$$

In the limit of large N, we have $q = \frac{v_2(s)}{v_1(s)+v_2(s)}$. For the case $f(x) \equiv g(x)$, we get a simple relation from (7):

$$q = \frac{\alpha_1}{\alpha_1+\alpha_2}. \tag{8}$$

This relation has been confirmed well by numerics (unpublished data).

The degenerated model with the uni-directional mutations of mutator gene.

The probability of back mutation (from mutant to wild type) is rather small and can be neglected in finite populations. Let us consider the infinite population model with $\alpha_2 = 0, \alpha_1 = a$ ($\mu_1 = 1, \mu_2 = \mu$). In this case, the expressions obtained for potential functions $V_\pm(x)$ are still valid, but it is necessary to take into account both branches of potential V_+ and V_- and define the mean fitness as their maxima.

Taking the $V_-(x)$ in Eq.(5), we get a new "mutator" phase (here and further the terminology is from [12]):

$$\begin{aligned} R &= \max_x [g(x) + \mu(\sqrt{1-x^2} - 1)], \\ g(s) &= R. \end{aligned} \tag{9}$$

The latter equation for the surplus is the same as for ordinary Crow-Kimura model [25]. It can be derived directly from the system (1), if we assume that the vast majority of population has a mutator allele, and omit the contribution of P_l in the second equation of Eq.(1). In the mutator phase the mean fitness is defined by the fitness landscape of mutator genome. Further we will consider the case when two fitness landscapes are identical, $f(x) = g(x)$.

Taking $V_+(x)$ in Eq.(5) under the condition $\alpha_2 = 0$ leads to the following expression for the mean fitness of mixed phase of Eq.(5) (see Methods):

$$\begin{aligned} R &= Max[f(x) - a + \sqrt{1-x^2} - 1], \\ g(s) &= R. \end{aligned} \tag{10}$$

The numerics supports well our analytical finding for the mean fitness, see Fig. 2. In the mixed phase the mean fitness is defined by the the wild-type fitness and mutation rate. However, numerical calculations for linear fitness landscapes $f(x) = kx$ or simple form quadratic fitness landscapes $f(x) = kx^2/2$ shows again that the vast majority of population has a mutator allele. It is worth pointing out that this result (the vast majority of population has a mutant type) is correct for any smooth fitness landscapes for the long enough genome length, see Eq. (16). We verify that for single peak fitness there is a finite fraction of wild-type sub-population at the infinite genome limit.

Two phases have been found for the discrete time Eigen model in [12]. This model can be exactly mapped into the Crow-Kimura model at the large N limit, when the mutation rate per gene is fixed. Fig. 2 gives the comparison

of our analytical results for the mean fitness R with the numerics. In [12] the mean fitness has been calculated for $f(x) = k(x-1), \quad k \gg 1$.

The single peak model.

Let us consider a fitness function with zero value for any argument except $l = 0$: $f(x_0) = J$. We have three possible phases in the steady state. We provide the following expressions for the mean fitness in the mutator phase R_{mu}, mixed phase R_{mix} and non-selective phase R_{ns}, see Fig. 2:

$$R_{mix} = J - 1 - a, \quad R_{mu} = J - \mu, \quad R_{ns} = 0. \tag{11}$$

In the mixed phase the fraction of the peak sequences with non-zero fitness in populations tends to 1 at the limit $N \to \infty$. We first calculate P_0 considering the equation for dP_0/dt in (1) and ignoring P_2 term. The fraction the wild type sequences in the l-th Hamming class in population is $P_l = P_0(1/J)^l$, therefore for the fraction of population with normal mutator gene we have $\sum_{l=0}^{N} P_l = \frac{P_0 J}{J-1}$. Considering analogously the equation for dQ_0/dt and ignoring Q_2 term, we obtain $Q_0 = P_0 \frac{a}{\mu-1-a}$. From the other hand, the definition of R gives: $P_0 + Q_0 = (J - a - 1)/J$. Thus we derive for the P_0 the following equation:

$$P_0 = \frac{(J-a-1)(\mu-a-1)}{J(\mu-1)} \tag{12}$$

This expression allows us to calculate the mutator allele probability q using the equivalence $\sum_{l=0}^{N} Q_l = 1 - \sum_{l=0}^{N} P_l$:

$$q = 1 - P_0 \frac{J}{J-1} \tag{13}$$

At $N = 5000$ the accuracy of our analytical result is about 0.1%, see Fig. 3.

The degenerated model with a smooth fitness landscape. As we mentioned before, we have a mixed and a mutator phases.In the mutator phase, the numerics gives that mutant sub-population dominates with $q = 1$ even for finite N, so we can ignore the first chain of transitions and reduce the system to a standard Crow-Kimura model with mutation rate μ_2 and fitness function $g(x)$.

Let us consider now the mixed phase. If $\alpha_1 = \alpha, \quad \alpha_2 \to 0$, then Eq.(8) gives $q \to 1$ for this case. Assuming an ansatz $P_l(t) \equiv P(x,t) = \exp(Nu(x,t))$, we obtain

$$R = f(s_1) - \alpha. \tag{14}$$

In the Methods section we calculate $(1-q)$ for the system with general smooth symmetric fitness function and parameters $\alpha, \quad \mu$. For linear fitness $f(x) = kx$ and small values of α and μ_1/μ_2, we obtain a simple expression

$$1 - q = \exp\left[\frac{-N\alpha_1^2}{2(\sqrt{k^2/\mu_1^2 + 1} - 1)\mu_1\mu_2}\right]. \tag{15}$$

This result is well supported by numerics, see Fig. 4 and Table II for $1 - q \ll 1$, when the exponent is a large negative number. In population genetics one uses the fitness difference for one mutation S instead of k, and U as a total number of mutations per generation. Using the discrete time Eigen model to Crow-Kimura model mapping, $U = \epsilon\mu_1, \ S = \epsilon 2k/N, \ h = \epsilon\alpha_1$, where ϵ is the discrete time step, and h is the transition probability to mutator allele during one generation we get a condition

$$\frac{N\alpha_1^2}{\mu_2 k} \equiv \frac{2h^2}{U\mu S} \gg 1. \tag{16}$$

Otherwise the numerics gives $q \ll 1$ instead of Eq.(15). Here μ is the ratio of two mutation rates, and h is the transition probability from normal to mutator allele per generation. We see that q decreases as an inverse of μ.

The linear fitness case can be solved exactly. We introduce the generation functions $Q(z) = \sum_l Q_l z^l, \quad P(z) = \sum_l P_l z^l$, and Eq.(1) gives a system of equations for the generation functions:

$$\begin{aligned} N(R - k + \alpha + (1-z))P &= (-2kz + 1 - z^2)P' \\ (R - k + \mu(1-z))Q - \alpha P &= (-2kz + \mu(1-z^2))Q' \end{aligned} \tag{17}$$

R can be calculated from the first equation, putting a constraint that $P(z)$ is a N order polynomial. Another constraint is the probability balance condition, $P(1) + Q(1) = 1$. Solving this system, we can calculate P_l, Q_l with a relative

accuracy $O(1/\sqrt{N})$. The generating function method has been applied before for the investigation of Crow-Kimura model [32].

The Eigen model with mutator gene and random fitness landscape.

Consider now the following system of equations for the $p_i, q_i, 0 \leq i < 2^N$.

$$\frac{dp_i(t)}{dt} = \sum_j p_j e^{-\alpha} Q_{ji} - p_i \sum_j r_i(p_i + q_i),$$
$$\frac{dq_i(t)}{dt} = \sum_j q_j \hat{Q}_{ji} + \sum_j p_j (1 - e^{-\alpha}) Q_{ji} - q_i \sum_j r_i(p_i + q_i). \quad (18)$$

where r_i is the fitness function, and Q_{ij} and $\hat{Q}_{ij}$ are the transition probabilities, $Q_{ij} = q^{L-d(j,i)}(1-q)^{d(j,i)}$; q is the probability of errorless replication per nucleotide for the p_i sequences. The diagonal terms of the mutation matrix are $Q_{ii} = q^L \equiv Q \equiv e^{-\gamma}$, where $\gamma = -N\ln(q) \approx N(1-q)$ is the parameter of mutation in the Eigen model. $\hat{Q}_{ij} = \hat{q}^{N-d(j,i)}(1-\hat{q})^{d(j,i)}$; $\hat{q}$ is the probability of errorless replication per nucleotide for the q_i sequences. $d(i,j)$ is the Hamming distance between two sequences i and j, the number of point mutations to get the sequence i from the sequence j.

For the single peak fitness function with $r_0 = A$ and $r_i = 1, i \geq 1$, we have a mean fitness similar to the case of CK model: $QAe^{-\alpha}$ for the mixed phase, $\hat{Q}A$ for the mutator phase and 0 for the non-selective phase.

The random fitness landscape is one of reasonable approximations to real biological data [33]. It is well realized that such a fitness landscape is almost equivalent to the single peak fitness landscape [24]. If we assume a log-normal distribution of Wrightian fitnesses r_i,

$$\rho(r_i) \sim (\exp[-N\frac{(\ln r_i)^2}{c^2}]) \quad (19)$$

then we have a maximal fitness as $A = c\sqrt{ln2}$ [24], while the vast majority have a fitness 1. Using the formulas for the single peak fitness case from the Appendix, we obtain

$$q = [1 - e^{-\alpha}][1 + \frac{1}{A(Qe^{-\alpha} - \hat{Q})}] \quad (20)$$

The steady state distribution is the same for the continuous time Eigen model (18), and for the discrete time Eigen model. Then we have $U = \gamma, h \approx \alpha$. For the $\alpha \ll 1, U \equiv N(1-q) \ll 1, (A-1) \equiv s \ll 1$ we obtain

$$q = \frac{\alpha}{\alpha + U} \quad (21)$$

Thus the result does not depend on the s, if we assumed the random character of the fitness landscape.

Taking $U = 2 * 10^{-4}, \alpha = 5 * 10^{-7}$ [34],[35], we obtain $q = 0.25\%$ for the E coli, consistent with the result of [36].

Discussion

In this work we investigated the mutator phenomenon in evolutionary process in stable environment for infinite population size. Before some results have been found for the microscopic model in case of uni-directional mutation (normal allele → mutator allele) with a linear Malthusian fitness in 1-dimensional fitness landscape and the mutator and mixed phase have been already considered theoretically, which is very important for us. We first calculated exactly the phase structure of the model for the case of general fitness function, the exact mean fitness and surpluses for any scheme of mutations between mutator and non-mutator alleles in d-dimensional fitness space (see the Methods), solved the important case of the random fitness landscape. We derived formulas for the mutator allele probability, exact at large genome length. This is highly non-trivial (non-perturbative via the value of the backward transition rate mutator allele to normal allele) mathematical problem: see Eqs.(8),(7),(15). Eq.(15) gives the probability of the mutator allele of the linear fitness function model for the case when the mutator take over the entire population, which has been observed experimentally [37, 38]. The solution for the general smooth fitness function is derived in the Methods. When the condition by Eq.(16) is broken, the normal allele takes over the population. Both situations (either normal or mutator alleles) have been observed experimentally [34]. As we mentioned, both allele can take over the population, depending on the genome length, see Eq. (16). Contrary to the condition Eq. (16), the mutator probability equals 1 in case of mutator phase, regardless the genome length. The linear fitness case allows an exact solution. In [12] has been derived an expression for the mutator allele probability (in case of linear fitness function)

when the mutators are minority. Our results are consistent with the result of [12], moreover, our method allows exact solution of the model in this case. We gave the formula for the mutator probability in case of random fitness landscapes. The latter formula simply decsribes the case of a small fraction of mutators in the population.

A careful investigation of the model (phase structure, order parameters), taken from biological literature is important, because the cancer is assumed to be essentially collective phenomenon, even with some collective intellect [39]. Therefore better to analyze the related biologically motivated models as accurately as possible, looking possible phases with order parameters. The investigation of cancer using only numerical or approximate investigation of models could be similar to the attempt to fix watches with hammer instead of using lens, as we see below. The theoretical results about error threshold [40] have a direct biomedical impact in case of both viruses and cancer, where recently has been suggested to use an error catastrophe as a new therapeutic strategy for the treatment of the solid cancer [5]. Our Fig 1. qualitatively represents the situation with different stages of cancer: the mixed phase is the early, non-aggressive version of tumor, while mutator is an aggressive version of the tumor, might be related with metastasis; we should transfer the tumor to the non-selective phase to eradiate the tumor according to the strategy suggested in [4]. According to our Fig. 1, there are 4 different situations. In case of mixed II sub-phase, we can push the tumor to the non-selective phase (the eventual goal of error-catastrophe therapy)increasing α_1, the mutation rate in the first part of the genome, responsible for the stability. In the mutator I sub-phase, we should increase the mutation rate μ to push the tumor to the non-selective phase. In mutator II and mixed I sub-phases we need to increase both versions of mutation rates. The Fig. 1 is identically the same for the random fitness case. We have qualitatively the same situation in case of other fitness landscapes.

We should investigate the finite population version of the mutator model, as the finite population sizes sometimes drastically affects the evolutionary dynamics [3]. The problem already attracted an attention [14], but it is reasonable first to solve the simpler case of infinite population limit, only later investigate the harder problem of finite population.

Methods

Ignoring $O(1/N)$ correction terms, using Eq. (4) and the formulas $P_{l\pm1} = v_1 e^{Nu\pm 2u'}$ and $Q_{l\pm1} = v_2 e^{Nu\pm 2u'}$ in Eq.(1), we get

$$v_1(-u'_t + f(x) - \alpha_1 + \mu_1(\frac{1+x}{2}e^{2u'} + \frac{1-x}{2}e^{-2u'} - 1)) + v_2\alpha_2 = 0,$$
$$v_1\alpha_1 + v_2(-u'_t + g(x) - \alpha_2 + \mu_2(\frac{1+x}{2}e^{2u'} + \frac{1-x}{2}e^{-2u'} - 1)) = 0. \quad (22)$$

Here we denoted $u' = \frac{\partial u(x,t)}{\partial x}$ and $q \equiv \frac{\partial u(x,t)}{\partial t} = u'_t$. We can consider Eq.(22) as a homogeneous system of equations for v_1 and v_2, and its determinant is 0 for the consistency of the equations. We find the u'_t from the latter condition and write an equation $u'_t + H_\pm(x,u') = 0$, identifying $H_\pm$ as a Hamiltonian. Then we derive the potential function as $V_\pm(x) = min[-H(x,p)]|_p$.

For the case $f(x) = g(x)$, we have the following expression for the Hamiltonian:

$$-H_\pm = f(x) - \alpha + \frac{1+\mu}{2}(-1 + \frac{1+x}{2}e^{2p} + \frac{1-x}{2}e^{-2p}) \pm \frac{1}{2}\sqrt{(1-\mu)^2(-1+\frac{1+x}{2}e^{2p} + \frac{1-x}{2}e^{-2p})^2 + 4\alpha^2}. \quad (23)$$

At the maximum point of distribution, we take $u' = 0$ and get Eq.(6) from zero determinant condition. Putting the found value of s into Eq.(22), we obtain Eq.(7).

Using an ansatz $P_l = \exp[Nu(x,t)]$, we get an equation:

$$R = u'_t = f(x) - \alpha + [\frac{1+x}{2}e^{2u'} + \frac{1-x}{2}e^{-2u'} - 1) \quad (24)$$

and second equation in Eq.(9).

The mean fitness for the d-dimensional fitness landscape and mutator gene.

Such model has been considered in [14], with two parts of genome: lethal mutations and deleterious plus advantageous mutations. The genome is formally fractured into d parts with the lengths Ny_i. For normal mutator allele the fitness is $\sum_i f(x_i)$ and in mutated case $\sum_i g(x_i)$, where $x_i = 1 - 2l_i/(Ny_i)$ and l_i is the number of $-$ spins (alleles) in the i-th part of genome. We have mutation rates μ_i for the upper chain and ν_i for the abnormal mutator allele case, see [31] for the multidimensional fitness model without mutator gene. Now our expression is modified, and we should

calculate mean fitness R as a maximum of the potential $V_{\pm}$, defined as

$$V_{\pm}(x) = \frac{\sum_i f_i(x_i) + g_i(x_i) - y_i(\mu_i + \nu_i + \alpha_1 + \alpha_2)}{2} + \sum_i y_i \frac{\mu_i + \nu_i}{2}\sqrt{1 - x_i^2} \pm \frac{1}{2}\sqrt{(A - \alpha_1 + \alpha_2)^2 + 4\alpha_1\alpha_2},$$

$$A = \sum_i f_i(x_i) - g_i(x_i) + y_i(\mu_i - \nu_i)(\sqrt{1 - x_i^2} - 1). \quad (25)$$

The model considered in [14] corresponds to the choice: $d = 2, f_1(1) = g_1(1) = 0$ and $f_1(x) = g_1(x) = -\infty$ for $-1 \le x < 1$, $f_2(x) = g_2(x) = k_2(1 - x)$, only in case of finite population.

The mutator probability

When $Q_l \sim P_l$, then we consider an equation for the Q_l:

$$RQ_l = \alpha \exp[Nu(x,t)] + Q_l(g(x_l) + \mu(Q_{l-1}\frac{N - l + 1}{N} + Q_{l+1}\frac{l + 1}{N} - 1). \quad (26)$$

We can write an approximate solution:

$$Q_l \sim \frac{\alpha \exp[Nu(x,t)]}{m}. \quad (27)$$

When $P_l \ll Q_l$, let us do an ansatz $Q_l = \exp[N\bar{u}(x,t)]$. We get the following equation for $\bar{u}$:

$$R = g(x) + \mu[\frac{1 + x}{2}e^{2\bar{u}'} + \frac{1 - x}{2}e^{-2\bar{u}'} - 1). \quad (28)$$

We took the solution (24) at the interval $[s_3, 1]$ and the solution (28) at the interval $[s_1, s_3]$. We glue two solutions, assuming the smoothness of the derivative $u'(s_3) = \bar{u}'(s_3)$, thus we obtain the following equation for s_3:

$$R + f(s_3) + \alpha = (R + f(s_3))/\mu. \quad (29)$$

Putting a condition $u(s_1) = 0$ (it is equivalent assuming $\sum_l P_l = 1$), we get the following expression for the $\sum_l Q_l \sim Q_{l_0}, l_0 = \frac{1+s_2}{2}N$:

$$1 - q \sim \exp[N(\int_{s_1}^{s_3} u'(x) - N\int_{s_2}^{s_3} \bar{u}'(x))]\frac{\alpha}{\mu}, \quad (30)$$

where $u', \bar{u}'$ are calculated using Eqs. (24) and (28).

The expression in the exponent in Eq.(30) is exact at the limit $N \to \infty$.

We can apply our analytical results to predict what allele takes over the population. The analysis of Table 2 reveals, that when $K \equiv \exp[N(\int_{s_1}^{s_3} u'(x) - N\int_{s_2}^{s_3} \bar{u}'(x))] \ll 1$, then mutator takes over. When $1 - K \ll 1$, normal allele takes over the population.

The case of small α

Now we have the following equations:

$$\begin{aligned} s_2 - s_1 &= \frac{\alpha}{f'(s_1)}, \\ G(s_3 - s_1) &= \frac{G}{\mu}(s_3 - s_2)u(s_1) - \frac{G(s_3 - s_1)^2}{2} = u(s_1) - \frac{G(s_3 - s_1)^2}{2\mu}, \\ G &= \frac{f'(s_1)}{s_1}. \end{aligned} \quad (31)$$

Thus

$$1 - q = \exp[\frac{-N\alpha^2}{2f'(s_1)s_1\mu}]. \quad (32)$$

For the linear fitness $f(x) = kx$ we get

$$1 - q = \exp[\frac{-N\alpha^2}{2(\sqrt{k^2 + 1} - 1)\mu}]. \quad (33)$$

This solution is correct for

$$s_1 \gg \frac{\alpha}{\mu k}. \tag{34}$$

The Eigen model version of the mutator model.

Consider the first chain of equation in the system (18). We can consider it as an equation for the Eigenvalue R. Compared with the ordinary Eigen model, the matrix of the linear system is multiplied by $e^{-\alpha}$. Thus we can directly use our results from [40], and write the expression for the mean fitness of the mixed phase for the fitness function $r_i = f(m)$:

$$R = \max_m [e^{-\alpha} f(m) e^{\gamma(\sqrt{1-m^2}-1)} \tag{35}$$

For the single peak fitness we obtain for the selective phase [40]

$$R = QAe^{-\alpha} \tag{36}$$

For the non-selective phase we simply have $R = 1$.

In the mutator phase we just ignore the first chain of equations in Eq.(18), and get R as the mean fitness of the Eigen model with the

$$R = \max_m [f(m) e^{\hat{\gamma}(\sqrt{1-m^2}-1)}] \tag{37}$$

Consider now the distribution of population in the single-peak fitness case.

Ignoring the back mutations for the equation for q_0, and putting the value of R from Eq. (33), we obtain

$$q_0 = \frac{1-e^{-\alpha}}{A(Qe^{-\alpha} - \hat{Q})} p_0 \tag{38}$$

As all the sequences besides p_0, q_0 have a fitness 1, we have an equation $(q_0 + p_0) = \frac{R}{A-1}$, therefore

$$p_0 = \frac{QAe^{-\alpha}}{(A-1)[1 + \frac{1-e^{-\alpha}}{A(Qe^{-\alpha}-\hat{Q})}]} \tag{39}$$

In the ordinary Eigen model

$$\frac{\sum_l p_l}{p_0} = \frac{A-1}{QA-1} \tag{40}$$

This equation is valid in our case, as all the equations in the first chain of (18) are derived from the ordinary model [40] just multiplying by $e^{-\alpha}$. Two equations give together Eq. (20) of the text.

For the discrete time Eigen model, we have the probabilities $p_i(n), q_i(n)$ and the following iteration equations:

$$\begin{aligned} p_i(n+1) &= \frac{\sum_j p_j(n) e^{-\alpha} Q_{ji}}{\sum_j r_i(p_i(n) + q_i(n))} \\ q_i(n+1) &= \frac{\sum_j q_j(n) \hat{Q}_{ji} + \sum_j p_j(n)(1-e^{-\alpha}) Q_{ji}}{\sum_j r_i(p_i(n) + q_i(n))} \end{aligned} \tag{41}$$

[1] L.A. Loeb, C.F. Springgate, N. Battula, Errors in DNA replication as a basis of malignant change. Cancer Res,**34**,2311–2321 (1974).

[2] Loeb, L. A., Loeb, K. R. & Anderson, J. P. Multiple mutations and cancer. Proc. Natl. Acad. Sci. USA **100**, 776-781(2003).

[3] Raynes Y., Gazzara, M. R. & Sniegowski, P. D. Contrasting dynamics of a mutator allele in aexual populations of different size. Evolution, **66**, 2329-2334 (2012).

[4] Fox, E. J. & Loeb, L. A. Lethal mutagenesis: targeting the mutator phenotype in cancer. *Semin Cancer Biol.* **20**, 353-359 (2010).

[5] Loeb, L. A. Human cancers express mutator phenotypes: origin, consequences and targeting. Nature review cancers **11**,450–457(2011).

[6] D. Hanahan & R. A. Weinberg. The hallmarks of cancer. Cell **100**, 57 (2000).

[7] C. P. Nowell. The clonal evolution of tumor cell populations. Science **194**, 23–28(1976).

[8] L. M. F. Merlo, J. W. Pepper, B. J. Reid and C. C. Maley, Cancer as an evolutionary and ecological process, Nature review cancers,**6**,924–935(2006).

[9] M. Greaves, C. C. Maley,Clonal evolution in cancer, Nature, **481**,306-311(2012)

[10] F. Taddei et al. Role of mutator alleles in adaptive evolution. Nature **387**, 700-702 (1997).

[11] D. Kessler and H. Levine. Mutator dynamics on a smooth evolutionary landscape. Phys. Rev. Lett.**80**: 2012 (1998).

[12] A. Nagar and K. Jain, Exact phase diagram of quasispecies model with a mutator rate modifier, Phys. Rev. Letters **102**, 038101 (2009).

[13] E. Baake, M. Baake, and H. Wagner. Quantum Chain is Equivalent to a Model of Biological Evolution. Phys. Rev. Lett. **78**, 559 (1997). .

[14] C. S. Wylie, C.-M. Ghim, D. Kessler and H. Levine, The Fixation Probability of Rare Mutators in Finite Asexual Populations, Genetics **181**: 1595–1612(2009).

[15] F. Michor, J. Liphardt, M. Ferrari and J. Widom. What does physics have to do with cancer? Nature Reviews Cancer, 11, 657-670 (2011). .

[16] R. V. Sole,Phase transitions in cancer, in d' Onofrio et al., eds New Challenges for Cancer Systems Biomedicine, A. Springer-Verlag, 2013.

[17] J. F. Crow and M. Kimura, *An Introduction to Population Genetics Theory* (Harper Row, NY, 1970).

[18] M. Eigen. Self organization of matter and the evolution biological macromolecules. Naturwissenschaften, **5**: 465–523(1971).

[19] M. Eigen, P. Shuster. The hypercycle: A principal of natural selforganization, Berlin-Heidelberg, Springer (1979).

[20] M. Eigen, J. McCascill and P. Schuster. The Molecular Quasi-Species. Advances in Chemical Physics, **75**:149-263 (1989).

[21] D. M. M. and D. S. Fisher, The balance between mutators and nonmutators in asexual populations Genetics **188**, 997-1014 (2011).

[22] E. Kussell, and M. Vucelja, Rep. Prog. Phys. **77**, 102602(2014).

[23] P. C. W. Davies and C H Lineweaver. Cancer tumors as Metazoa 1.0: tapping genes of ancient ancestors. Phys. Biol. **8**, 015001 (2011).

[24] Z. Kirakosyan, D. B. Saakian, and C.-K. Hu. Evolution models with lethal mutations on symmetric or random fitness landscapes. Phys. Rev. E **82**, 011904 (2010).

[25] E. Baake and H. Wagner, Mutation-selection models solved exactly with methods of statistical mechanics, Genet. Res. **78**, 93–117 (2001).

[26] D. B. Saakian A new method for the solution of models of biological evolution: Derivation of exact steady-state distributions *Journal of Stat. Physics*, **128**:781(2007).

[27] K. Sato, K. Kaneko Evolution equation of phenotype distribution: General formulation and application to error catastrophe. *Phys. Rev. E* **75**,061909(2007).

[28] D. B. Saakian, O. Rozanova, A. Akmetzhanov Exactly solvable dynamics of the Eigen and the Crow-Kimura model. *Phys. Rev. E* **78**: 041908 (2008).

[29] G. Woodcock and P. G. Higgs,Population evolution on a multiplicative single-peak fitness , J. Theor. Biol. **179**: 61-7-3 (1996).

[30] C. J. Thompson and J. L. McBride,On The Eigen's theory of selforganization of matter, Math. Biosci. **21**, 127 (1974).

[31] D.B. Saakian, V. Galstyan, Dynamics of the Chemical Master Equation, a strip of chains of equations in d-dimensional space, submitted to Phys. Rev. E. **86**, 011125 (2012).

[32] A.S. Bratus, A. S. Novozhilov, Y.S. Semenov, Linear algebra of the permutation invariant Crow-Kimura model of prebiotic evolution, Mathematical Biosciences, **256**:42–57(2014).

[33] R. Sanjuan, A. Moya, and S. F. Elena,The distribution of fitness effects caused by single-nucleotide substitutions in an RNA virus, Proc. Natl. Acad. Sci. U.S.A. 101, 8396 (2004).

[34] L. Boe, M. Danielsen, S. Knudsen, J. B. Petersen, J. Maymann, P. R. Jensen, The frequency of mutators in populations of Escherichia coli, Mutation Research **448**,47-55(2000).

[35] J. Ninio,Transient Mutators: A Semiquantitative Analysis of the Influence of Translation and Transcription Errors on Mutation Rates, Genetics 129, 957 (1991).

[36] M. D. Gross and E. C. Siegel,Proliferation of mutators in A cell population. Mutat. Res.**91**, 107–110 (1981).

[37] P. D. Sniegowski, P. J. Gerrish, R. E. Lenski, Evolution of high mutation rates in experimental populations of E. coli, Nature **387**,703-705(1997).

[38] L. Chao, E.C. Cox, Competition between high and low mutating strains of Escherichia coli, Evolution **37**,125-134(1983).

[39] M. Tarabichi et al Systems biology of cancer: entropy, disorder, and selection-driven evolution to independence, invasion and swarm intelligence, Cancer Metastasis Rev **32**:4031(2013).

[40] D. B. Saakian and C.-K. Hu, Exact solution of the Eigen model with general fitness functions and degradation rates, Proc. Natl. Acad. Sci. USA Proc. Natl. Acad. Sci. U.S.A. **103**, 4935 (2006).

Acknowledgments
DBS thanks R. Sole for the discussion. DBS and CKH thank Taiwan-Russia collaborative research grant with Grant Number 101-2923-M-001 -003 -MY3, NCTS (North), and Academia Sinica for support.

Author contributions
DBS and CKH designed and performed the research as well as wrote the paper. YT did numerical calculations.

Additional information
Competing financial interests: The authors declare no competing financial interests.

TABLE I: The comparison of the results for $f(x) = g(x) = 3x^2/2, \mu_1 = 1, \mu_2 = \mu, \alpha_1 = \alpha_2 = 1, N = 400$. R_{num} is the numerical result and R_{th} is given by Eq.(5).

μ	3.5	3.	2.5	2.	1.5	1.	0.5
R_{num}	0.1907	0.2514	0.32405	0.4127	0.5240	0.6684	0.8626
R_{th}	0.1811	0.2436	0.3180	0.4084	0.5212	0.6666	0.8615

TABLE II: The results for $f(x) = g(x) = kx, \mu_1 = 1, \mu_2 = \mu, \alpha_1 = \alpha, \alpha_2 = 0$. $K = \exp[N(\int_{s_1}^{s_3} u'(x) - N\int_{s_2}^{s_3} \bar{u}'(x))]$. R_n is a numerical result.

N	1000	1000	1000	1000
k	0.3	0.3	0.3	1
α	0.0001	0.001	0.01	0.3
R_n	0.0439	0.0430	0.0340	0.1142
R	0.0439	0.0430	0.0340	0.1142
q	0.9945	0.9460	0.530	$6/10^7$
K	1-$4/10^6$	0.9994	0.930	$1/10^5$

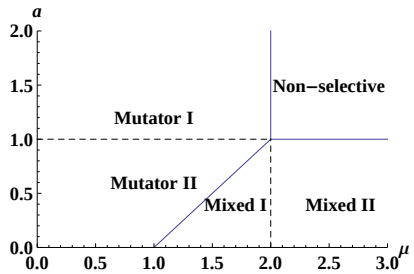

FIG. 2: The phase structure of the mutator model with single peak landscape, $\alpha_1 = a, \alpha_2 = 0, \mu_1 = 1, \mu_2 = \mu$. There are three phases: mixed phase with $0 < s < 1, 0 < q < 1$, non-selective phase with $s = 0, 0 < q \leq 1$ and mutator phase with $0 < s, q = 1$. The border between non-selective and mutator phases is given by $\mu = J$, the border between non-selective and mixed phases is given by $a = J - 1$, between mixed and mutator phases is given by $a + 1 = \mu$ line. From the bio-medical perspectives we distinguish the mutator I and mutator II, mixed I and mixed II sub-phases. From the mutator I, the system transforms to the non-selective phase simply increasing the μ. From the mixed II the system transformers to the non-selective phase simply increasing the $a \equiv \alpha_1$. From the mutator II and mixed I sub-phases we need change both a and μ to transform the system to the non-selective phase.

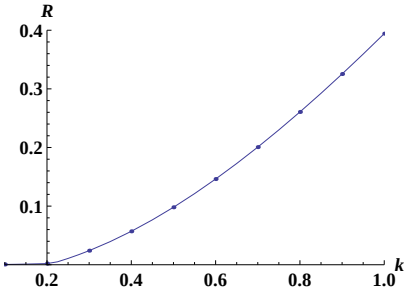

FIG. 3: The mean fitness R versus k of the model with linear fitness landscape $f(x) = kx$, $\mu_1 = 1$, $\mu_2 = 10$, $\alpha_2 = 0$, $\alpha_1 = 0.02$. There are two phases in the model for the general values of parameters: mixed phase with $R = \sqrt{k^2+1} - \alpha_1 - 1$ and mutator phase with $R = \sqrt{k^2 + (\mu_2)^2} - \mu_2, q = 1$. The border between two phases is given by equation $\alpha_1 = \mu_2 - 1 + \sqrt{k^2+1} - \sqrt{k^2 + (\mu_2)^2}$. In our case $k_c \approx 0.212$.

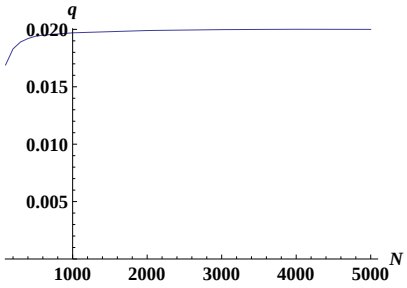

FIG. 4: The dependence of the mutator probability q on the genome length N. The the single peak fitness model (smooth line) with $J = 1.05, \mu_1 = 1, \mu_2 = 10, \alpha_1 = 0.001$. For the $N = 5000$ the single peak model's numerical result coincides with the analytical result for $N = \infty$ with the relative accuracy about 0.1%.