Fluctuation theorem for the mutation process in *in vitro* evolution*

Liu Qi(刘 琪)[a)b)], Tang Chao(汤 超)[b)c)], and Ouyang Qi(欧阳颀)[a)b)†]

[a)] *Department of Physics, Peking University, Beijing* 100871, *China*
[b)] *Center for Theoretical Biology, Peking University, Beijing* 100871, *China*
[c)] *California Institute for Quantitative Biomedical Research, Departments of Biopharmaceutical Sciences and Biochemistry and Biophysics, University of California, San Francisco, CA* 94143, *USA*

(Received 29 June 2009; revised manuscript received 17 September 2009)

A proposition based on the fluctuation theorem in thermodynamics is formulated to quantitatively describe molecular evolution processes in biology. Although we cannot give full proof of its generality, we demonstrate via computer simulation its applicability in an example of DNA *in vitro* evolution. According to this theorem, the evolution process is a series of exponentially rare fluctuations fixed by the force of natural selection.

Keywords: fluctuation theorem, molecular evolution, natural selection

PACC: 0200, 0540, 8710, 8725

1. Introduction

The second law of thermodynamics states that all naturally occurring processes in closed systems evolve irreversibly in one "time-forward" direction. However, how reversible microscopic equations of motion can lead to irreversible macroscopic behaviour was not quantitatively understood, until the discovery of fluctuation theorem in the last decade.[1] This theorem not only proves that in a large system observed for a long period of time the second law is overwhelmingly likely to be valid, but also quantifies the probability of observing the deviation from the second law, especially in small systems. Since then, several related fluctuation theorems were deduced;[2−4] they dramatically expanded the applicability of the original theorem. The generality of the theorem was summarized in Ref. [5], and a number of experiments and a few modellings were conducted to test its validity in a variety of disparate circumstances.[6−15] Interestingly, the theorem was corroborated even in open systems, and to local variables by experiments,[16,17] which is beyond the prediction of the theorem. Also, the theorem was expanded to chemical reactions,[18,19] further demonstrating its generality. In this paper, we propose that the fluctuation theorem may also be used to quantitatively describe a more complex system—evolution.

According to Darwin's theory of evolution, life evolves forward by random mutation and subsequent natural selection in a "survival of the fittest" scenario. The mapping between the genotype where the mutation takes place and the phenotype which the selection acts upon is usually very complicated, so it is difficult to know the probability that a generation will improve its fitness (or adaptability) as a result of beneficial mutations. However, under the assumption that life evolves towards more complex states (so the genome is evolving to more "ordered" structures), there seems to be an analogy between "fluctuations" in physical systems and mutations in the evolutionary process. Here we show in a specific example of *in vitro* DNA evolution that there exists a quantitative relationship between the amounts of increasing and decreasing fitness as a result of random mutations that is very similar to the fluctuation theorem.

2. Theoretical formulation

To formulate a theory of fluctuation in evolution, we define α as a measure of the adaptability (or fitness) of an individual. It will be a function of a variety of variables in the real world, thus $\alpha = \alpha(x_1, x_2, x_3, \ldots)$. Yet, how to define such a parameter is somewhat arbitrary and ambiguous. As

*Project supported by the National Natural Science Foundation of China (Grant No. 10721403), the National Basic Research Program of China (Grant No. 2007CB814802), and the Jun-Zheng Foundation at Peking University.
†Corresponding author. E-mail: qi@pku.edu.cn
© 2010 Chinese Physical Society and IOP Publishing Ltd

long as it can depict the fact that the larger the value of α, the more adaptable the individual, it might be regarded as a reasonable order parameter.

Next, we concentrate on the evolution of a certain species. As is pointed out in Darwin's evolution theory, the process of mutation is of vital importance in evolution. Whenever an individual produces its offspring, the chance is high that the adaptability of its offspring will decrease through mutation. Only a small portion of all offspring will become more adaptable to the environment, and nature selects these ones to drive the species' evolution forward. In other words, under the driving of selection pressure, the value of α increases. To further study the change of adaptability in evolution, we define a function $f(\alpha)$ to represent the natural probability of an individual having the value of α. Suppose the parameters $x_1, x_2, x_3, \ldots$ in $\alpha(x_1, x_2, x_3, \ldots)$ are randomly distributed, for the reason that the state of larger α is a more ordered state and is harder to obtain, $f(\alpha)$ should be a monotonically decreasing function of α. From this point of view, an evolution process is a trajectory on the $f(\alpha) \sim \alpha$ curve, as shown in Fig. 1.

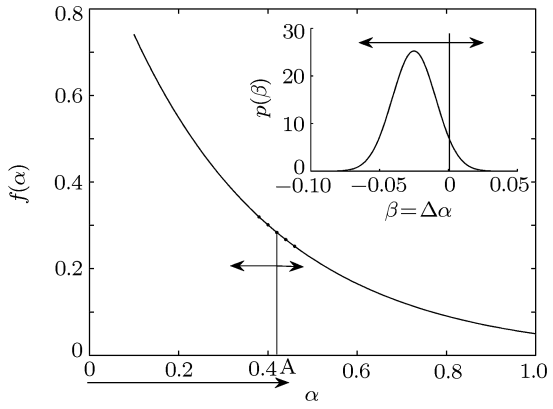


Fig. 1. Under the driving of selection pressure, α increases gradually. Considering an individual at point A, when it produces offspring with random mutations, most of them would become less adaptable. Inset figure shows a typical distribution of $\beta = \Delta\alpha$. A small portion of descendants would have positive changes in α and would evolve forward by natural selection.

In Fig. 1, we assume that an individual is at point A, and consider the adaptability of its offspring. For brevity, we define $\beta = \Delta\alpha$ to measure the change of the adaptability between the two generations. The probability that $\beta > 0$ will be quite small compared to the probability that $\beta < 0$. This probability distribution might be different in different circumstances. However, we show in the following that if β is small, the equality will hold

$$\frac{P(\beta)}{P(-\beta)} = \mathrm{e}^{-k\beta}, \tag{1}$$

where $P(\beta)$ is the probability that α will increase β, $P(-\beta)$ is the probability that α will decrease β, and k is a constant which is different in different situations.

We use a biased random walk to prove the above equation. First we discretize the axis of α with increment δ, which is infinitesimally small compared to β. The states of offspring can be determined from a random walk along the α axis. The probability that a point jumps a step forward (P_{F}) compared to the probability that the point jumps backward (P_{B}) can be determined from the function of $f(\alpha)$:

$$\frac{P_{\mathrm{F}}}{P_{\mathrm{B}}} \approx \frac{f(\alpha+\delta)}{f(\alpha)} \approx \frac{f(\alpha)}{f(\alpha-\delta)} \approx 1 + \frac{f'(\alpha)}{f(\alpha)}\delta, \tag{2}$$

where $f'(\alpha) < 0$. Since β is quite a small number, the ratio of $f'(\alpha)/f(\alpha)$ is almost constant in the first approximation. Assuming that point A walks N steps in total, the probability that it walks n steps forward is

$$P(+n) = \binom{n_+}{N} \left(\frac{P_F}{P_F + P_{\mathrm{B}}}\right)^{n_+} \left(\frac{P_{\mathrm{B}}}{P_F + P_{\mathrm{B}}}\right)^{n_-},$$

where $n_+ = (N+n)/2$, $n_- = (N-n)/2$. Likewise, the probability that it walks n steps backward is

$$P(-n) = \binom{n_-}{N} \left(\frac{P_{\mathrm{B}}}{P_F + P_{\mathrm{B}}}\right)^{n_+} \left(\frac{P_F}{P_F + P_{\mathrm{B}}}\right)^{n_-}.$$

Consequently, the ratio is

$$\frac{P(+n)}{P(-n)} = \left(\frac{P_F}{P_{\mathrm{B}}}\right)^n = \exp\left(-n\ln\left(\frac{P_{\mathrm{B}}}{P_F}\right)\right).$$

Defining $k \equiv \ln(P_{\mathrm{B}}/P_{\mathrm{F}})/\delta \approx -f'(\alpha)/f(\alpha) > 0$, and noting that $n\delta = \Delta\alpha = \beta$, we come to Eq. (1). Using $f'(\alpha) < 0$, it can also be deduced that the average change of α is negative, that is, $\overline{\beta} = N\delta(P_{\mathrm{F}} - P_{\mathrm{B}}) = N\delta P_{\mathrm{B}} f'(\alpha)/f(\alpha) < 0$. A typical distribution of β is shown in the inset of Fig. 1.

If we hope Eq. (1) is valid even if β is not small, the right-hand side of Eq. (2) should be a constant and independent of α. In this case, one can easily deduce that $f(\alpha) = f_0 \mathrm{e}^{-k'\alpha}$.

A generalised version of the fluctuation theorem is the entropy production theorem, which is valid for systems driven arbitrarily far from thermodynamic equilibrium. Although the types of systems, range of applicability, and exact interpretation differ, this theorem has the following general form[5]

$$\frac{P[\Omega(t) - \Omega(0) = A]}{P[\Omega(t) - \Omega(0) = -A]} = \exp(A), \tag{3}$$

where $\Omega(t)$ is the amount of dissipation in the system. Without losing generality, by considering a physical Hamiltonian system composed of N particles in a box, and applying an external Hamiltonian $H'(x)$, it will come to equilibrium with a Boltzmann distribution

$$\pi_d(x) = A'\mathrm{e}^{-[H_0+H'(x)]/k_\mathrm{B}T} = A\mathrm{e}^{H'(x)/k_\mathrm{B}T}, \quad (4)$$

where k_B is the Boltzmann constant and T the temperature. If we turn off the external field $H'(x)$, we can expect that the system will go back to its original equilibrium state, and the average entropy production will be positive. However, it can be proved[5] that if we define $\Omega(t) \equiv \sum_{i=1}^{N} H'(x_i(t))/k_\mathrm{B}T$, the distribution of $\Omega(t)$ satisfies Eq. (3) at any time.

Now supposing that the adaptability change of a species during one generation of evolution can be considered as the negative entropy production ($\beta \sim -A$), equation (3) is equivalent to the fluctuation theorem of Eq. (3) by scaling β with a factor k. Thus, the evolution process can be viewed as some kind of fluctuation process that satisfies the fluctuation theorem. The natural selection is then analogous to an external field that favours a certain direction of the fluctuation. In the following, we demonstrate that the mutation process in DNA *in vitro* evolution obeys this equation.

3. Simulation of DNA *in vitro* evolution

In vitro evolution has been used as a testing ground for evolution theory because of its ease of quantisation. It also has practical applications in molecular biology. Recently much attention has been paid to both experiment and theory.[20−23] Here we focus our attention on the process of evolving a pool of DNA sequences towards better binding to a protein.

We use a model similar to that of Yang *et al.*[20] to study the DNA *in vitro* evolution process. The goal of the process is to evolve a sequence of DNA that specifically binds to a target protein, the *Mnt*-repressor. In the model, a population of N random DNA sequences, each consisting of L nucleotides is prepared at the beginning. The sequences are subject to several cycles of duplication, with an error rate ν per nucleotide per duplication. After M cycles of duplication, the population size becomes $N \cdot 2^M$. In the subsequent selection process, N pieces of DNA are selected according to the binding probability to the protein. The duplication process followed by the selection process is called one generation of the evolution. The evolution of the DNA sequences is carried out through many generations. To understand the selection process, we consider the equilibrium reaction of binding-unbinding between a DNA sequence and a protein: $S + MR \rightleftharpoons S - MR$, where $S = b_1 b_2 \ldots b_L$ is a DNA sequence of nucleotides b_i, which can be A, C, G, and T, MR represents the *Mnt*-repressor, and S–MR is the DNA-protein complex. According to this model, the binding probability of sequence S by *Mnt*-repressor molecules has the following form:

$$P(S,\mu) = \frac{1}{1+\exp\left(\dfrac{-E_S-\mu}{k_\mathrm{B}T}\right)}, \quad (5)$$

where μ is the chemical potential that can be controlled by the number of available proteins and the choice of binding buffer. E_S is the binding energy of S to *Mnt*-repressor, which consists of two parts: specific binding energy ϵ_s and nonspecific binding energy ϵ_0 that does not depend on specific sequences. Under the assumption[24] that each nucleotide in the sequence independently contributes to the specific binding energy ϵ_s, one has $\epsilon_\mathrm{s} = \sum_{i=1}^{L} \epsilon_{b_i}$, where ϵ_{b_i} is the energy contribution of the nucleotide b_i in the sequence S. In our simulation, ϵ_{b_i} are taken from the experimentally measured values in Ref. [24]. Therefore, for any given DNA sequence S, $P(S,\mu)$ can be determined from Eq. (5).

If N is much larger than the number of all possible DNA sequences of length L (4^L) and the effect of mutation can be neglected, the state of the DNA pool at time t (in units of generations) can be described with the sequence distribution function:[20]

$$\pi_d(S,t) = \frac{\pi_d(S,0)}{\sum_S P^t(S,\mu)\pi_d(S,0)}[P(S,\mu)]^t, \quad (6)$$

where $\pi_d(S,0) = 1/4^L$. Defining $\alpha' \equiv \ln P(S,\mu)$, we write Eq. (6) as follows:

$$\pi_d(\alpha',t) = A\mathrm{e}^{t\alpha'}. \quad (7)$$

We define an order parameter α to measure the adaptability of the DNA pool: $\alpha \equiv \sum_{i=1}^{N} \alpha'(S_i)$, where each pool is regarded as an individual with an adaptability α'. Equation (7) governs the selection of the evolution process. The adaptability α will increase with time, and the distribution will move forward, as shown in Fig. 2.

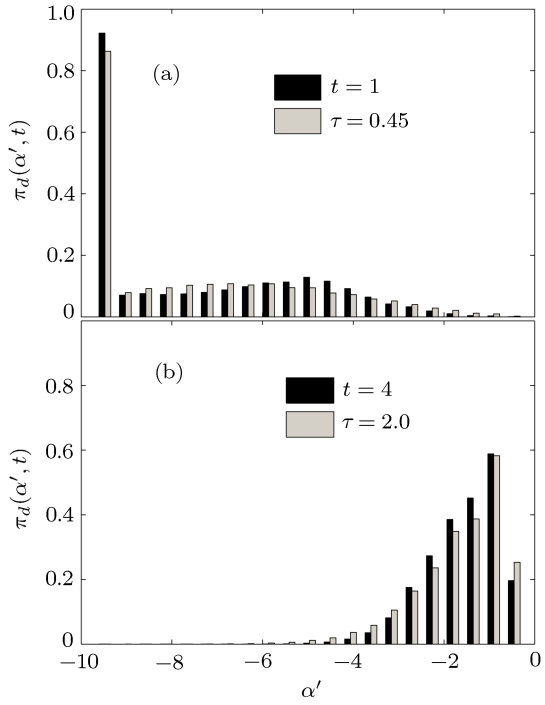


Fig. 2. The distribution of α' in the first (a) and the fourth (b) round of evolution. The distribution moves forward as time increases. Black boxes show the simulation results, grey boxes present the fitting of Eq. (7) with an effective time τ. (a), $t = 1$, $\tau = 0.45$; (b), $t = 4$, $\tau = 2.0$. The other parameters in simulation are: $N = 500000$, $M = 3$, $L = 9$, $\nu = 0.01$, $\mu = 0.5k_{\rm B}T$.

However, because equations (6) and (7) do not include mutations, they do not hold for large mutation rates. Nevertheless, we find that if we substitute time t with an effective time τ and regard the latter as a free parameter, we can always find a suitable value to fit the distribution. The grey bars in Fig. 2 show some examples of the fitting. Therefore, rather than discarding Eq. (7) totally, we replace t in Eqs. (6) and (7) with a corresponding parameter τ to approximate the distribution of DNA pool with mutations. This approach gives an analytical description of the distribution of the DNA pool in each round of the *in vitro* evolution. Table 1 lists the estimated effective time τ in each generation.

Table 1. The estimated effective time τ and the slope k of the lines (Fig. 3) for each round of evolution t.

t	1	2	3	4	5	6	7	8
τ	0.45	1.0	1.4	2.0	2.6	3.0	3.2	3.5
k	0.65	1.21	1.37	1.43	1.55	1.78	1.84	1.90

The rest of the problem is to show that the change in α during the mutation process follows Eq. (1). We use computer simulation to test the proposed theory. At a stage in the middle of the evolution process, we duplicate the evolved pool many times, introduce a mutation rate ν in each pool, and measure the distribution of the changes of α. We define $\beta = \Delta\alpha$, and our simulation results (Fig. 3(a)) show that the probability density function $P(\beta)$ decays exponentially when $|\beta|$ is large and that both of the most probable values β and $\overline{\beta}$ are negative, so without the selection process, the DNA pool will eventually evolve back to a random distribution. To test Eq. (1), we calculate the ratio of $P(\beta)/P(-\beta)$. Figure 3(b) shows that the value of $-\ln[P(\beta)/P(-\beta)]$ is a linear function of β, in excellent agreement with the theory. Notice that because of Eq. (7), equation (1) holds even for large values of β in the mutation process in *in vitro* evolution.

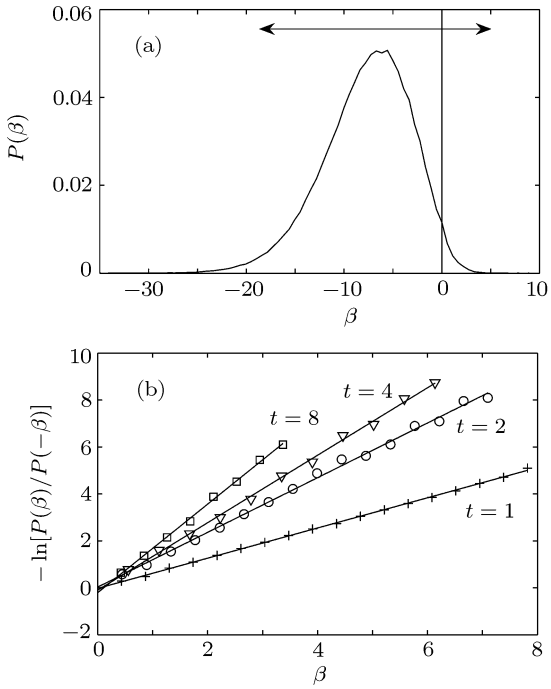


Fig. 3. Computer simulations to calculate $P(\beta)$ in the middle of DNA *in vitro* evolution. (a) Probability density of β at time $t = 2$. Note that there is only a very small possibility for $\alpha > 0$. (b) The function $-\ln[P(\beta)/P(-\beta)]$ versus β for different time $t = 1$, 2, 4, 8. The parameters in simulation are: $N = 100$, $M = 3$, $L = 9$, $\nu = 0.01$, $\mu = 0.5k_{\rm B}T$.

It is instructive to calculate the value of k theoretically, and compare it with the slope of the line drawn in Fig. 3(b). In fact, for a given distribution of N sequences of DNA according to Eq. (7), it is apparent that it will go back to the random distribution through mutation. This is quite similar to the relaxation process in the above-mentioned Hamiltonian system. If we regard the driving of natural selection as the external Hamiltonian $H'(x)$, i.e., $\tau\alpha' \sim H'(x)/k_{\rm B}T$, equa-

tions (7) and (4) become identical. Therefore, we can deduce Eq. (1) from Eq. (3), with the scaling factor τ: $\Omega \equiv \sum H'(x_i)/k_{\rm B}T \sim \tau \sum \alpha' \equiv \tau\alpha$. In other words, we expect that the slope of the line in Fig. 3(b) is $k = \tau$.

In Table 1 we compare the value of τ determined from the distribution of α' with the value of k measured from the slope of Fig. 3(b). It shows that τ is much closer to the slope of the line than time t. However, k is still smaller than the effective time τ. The reason is that for the sake of saving simulation time, in the simulation we choose N to be a relatively small number compared with 4^L.[1)] The small number N contributes to the inaccuracy of Eq. (6), which induces a deviation between theoretical and simulation results. Yet the perfect lines in Fig. 3 demonstrate the generality of the fluctuation theorem to some extent.

4. Discussion

The central idea of this paper is to draw an analogy between thermodynamic fluctuation in equilibrium systems and mutation in Darwinian evolution and to propose that the same form of the fluctuation theorem (FT) applies. In equilibrium systems the FT quantifies the probability of fluctuations that deviate from the second law; in the evolution process the FT quantifies the probability of mutations that benefit life. The fundamental difference is that while in equilibrium systems these rare fluctuations are transient, in the evolution process they can be fixed by selection so that the system is poised for further evolution. We note that the real evolution process is much more complicated than described here. For example, besides random point mutations, nature invents other ways to increase the chances of beneficial "fluctuations" and to buffer the bad ones, including gene duplication, recombination, genome reorganisation, etc. Also, the environment is constantly changing and there is no such thing as a fixed fitness landscape and thus a fixed "ideal" genome to evolve towards. However, we expect that our work will shed light on a more quantitative understanding of the evolution process at least in some special cases and we hope that it will stimulate further research in this direction.

References

[1] Evans D J, Cohen E G D, and Morriss G P 1993 *Phys. Rev. Lett.* **71** 2401

[2] Jarzynski C 1997 *Phys. Rev. Lett.* **78** 2690

[3] Crooks G E 1999 *Phys. Rev.* E **60** 2721

[4] Crooks G E 2000 *Phys. Rev.* E **61** 2361

[5] Evans D J and Searles D J 2002 *Advance in Physics* **51** 1529

[6] Wang G M, Carberry J C, Williams D R M, Sevick E M and Evans D J 2005 *Phys. Rev.* E **71** 046142

[7] Schuler S, Speck T, Tietz C, Wrachtrup J and Seifert U 2005 *Phys. Rev. Lett.* **94** 180602

[8] Wang G M, Sevick E M, Emil Mittag, Searles D J and Evans D J 2002 *Phys. Rev. Lett.* **89** 050601

[9] Collin D, Ritort F, Jarzynski C, Smith S B, Tinoco I Jr and Bustamante C 2005 *Nature* **437** 231

[10] Trepagnier E H, Jarzynski C, Ritort F, Crooks G E, Bustamante C J and Liphardt J 2004 *PNAS* **101** 15038

[11] Liphardt J, Dumont S, Smith S B, Tinoco I Jr and Bustamante C 2002 *Science* **296** 1832

[12] Carberry D M, Reid J C, Wang G M, Sevick E M, Searles D J and Evans D J 2004 *Phys. Rev. Lett.* **92** 140601

[13] Chen S B and Mei D C 2006 *Chin. Phys.* **15** 2861

[14] Cao L, Cheng Q H, Wu D J and Xu D H 2006 *Chin. Phys.* **15** 2324

[15] Chen Y D, Hu J M, Li L and Zhang Y 2009 *Chin. Phys.* B **18** 1373

[16] Shang X D, Tong P and Xia K Q 2005 *Phys. Rev.* E **72** 015301

[17] Ciliberto S, Garnier N, Hernandez S, Lacpatia C, Pinton J F and Chavarria G R 2004 *Physica* A **340** 240

[18] Seifert U 2004 *J. Phys.* A: *Math. Gen.* **37** L517

[19] Gaspard P 2004 *J. Chem. Phys.* **120** 8898

[20] Yang Y, Wang H L and Ouyang Q 2003 *Phys. Rev.* E **68** 031903

[21] Dubertret B, Liu S, Ouyang Q and Libchaber A 2001 *Phys. Rev. Lett.* **86** 6022

[22] Peng W, Gerland U, Hwa T and Levine H 2003 *Phys. Rev. Lett.* **90** 088103

[23] Kloster M and Tang C 2004 *Phys. Rev. Lett.* **92** 038101

[24] Fields D S, He Y Y, Al-Uzri A Y and Stormo G D 1997 *J. Mol. Biol.* **271** 178

[1)] We find that the larger N is, the less the probability of $\beta > 0$. This means that the area under $P(\beta)$ in the positive axis of β is smaller. To acquire enough points to fit the line in Fig. 3(b), we require that the probability of $\beta > 0$ should not be too small.