Dynamics of DNA *in vitro* evolution with *Mnt*-repressor: Simulations and analysis

Yufeng Yang,[1,2] Hongli Wang,[1] and Qi Ouyang[1,2,*]
[1]*Department of Physics, Peking University, 100871 Beijing, China*
[2]*Center for Theoretical Biology, Peking University, Beijing 100871, China*
(Received 21 February 2003; revised manuscript received 13 May 2003; published 9 September 2003)

The dynamics of DNA *in vitro* evolution with *Mnt*-repressor has been studied numerically and analytically. Based on experimental data and realistic energy landscape for DNA-*Mnt*-repressor interaction, we investigated the dynamics of DNA *in vitro* evolution using stochastic simulations. The binding energy of DNA to *Mnt*-repressor was considered to consist of two parts: the DNA sequence specific and nonspecific. The crossover observed in real experiments is numerically recovered. We demonstrate that the evolution trajectories are drastically dispersed and no typical evolution passage exists during the evolution. Particularly, Fisher's theorem of natural selection is verified. A theoretical analysis for the evolution is also included.

DOI: 10.1103/PhysRevE.68.031903 PACS number(s): 87.14.Gg, 87.15.Aa, 87.23.Kg, 02.70.Uu

I. INTRODUCTION

In vitro evolution is becoming an important tool in molecular biology. It is widely used to develop novel proteins, DNA, or RNA for special purposes [1–5]. Schematically, an *in vitro* evolution is a process of repeated operation cycles within which reproduction, mutation, and selection processes take place consecutively. In the case of DNA *in vitro* evolution, the reproduction and mutation can be realized by the polymerase chain reaction (PCR), and the selection can be a binding process where DNA sequences are selected according to their respective binding affinity to proteins [6] or other kinds of molecules. In 1990s, a relevant mathematics was developed to provide analytical insight into the *in vitro* enrichment process [7,8]. Recently, the dynamics of competitive DNA *in vitro* evolution via protein binding has been analytically studied by Peng *et al.* [9]. Based on a continuum mean-field model, they conclude that interactions between mutations and the selection pressure can drive the system to an asymptotic equilibrium state where the population distribution centers at a sequence which can be far away from the best sequence that the protein binds. The first quantitative experimental study on the dynamics of DNA *in vitro* evolution was carried out by Dubertret *et al.* [10]. They studied the evolution of a random pool of DNA sequences under the selection pressure of *lac*-repressor and discovered several significant dynamical features. One of which is the convergence of evolving sequences to the best protein-binding sequence, which has an abrupt change during the evolution. At present, the phenomenon has not been explained.

In this paper, we report numerical and analytical studies of the dynamics of DNA *in vitro* evolution that occur in real situations. In the simplified model of Ref. [9], the basic hypothesis is that the DNA-protein-binding energy is determined simply by the number of nucleotides in the DNA that are distinct from the best sequence. We here use a more realistic assumption that takes into account practical features: the binding energy of DNA sequences to proteins consists of two parts: the sequence-specific energy and the sequence-nonspecific energy. The former part is obtained from the experimental data [11] and the latter is a parameter in our model. Our simulation also tries to mimic the experimental procedure of DNA *in vitro* evolution [10]. In each evolution cycle, a population is amplified by a number of PCR cycles and diversified simultaneously due to nucleotide mutations. The amplified population is then introduced to a tube of binding buffer where the protein is fixed on the wall of the tube. A fraction of DNAs is bound to the protein; they are then separated and released from the tube. This population of DNA is used as the input for the next evolution cycle. We investigated the dynamics of DNA *in vitro* evolution by carrying out stochastic simulation. The crossover process observed in real experiments is numerically recovered. We demonstrate numerically that the evolution trajectories are drastically dispersed during evolution and there does not exist any typical evolution pathway. Particularly, Fisher's theorem of natural selection is verified in our simulations. A simplification for the experimental data is also discussed theoretically in the context of our model.

II. MODEL

In our numerical model, a population of N random DNA sequences, each consisting of L nucleotides is first prepared. The sequences are then subject to I cycles of duplication, with a small error rate ν_0 (typically in a magnitude of 10^{-4}) per nucleotide when duplicated. The selection is fulfilled through an equilibrium reaction: $S+\mathrm{MR}\rightleftharpoons S\text{-MR}$, where $S=b_1b_2\cdots b_L$ is a DNA sequence of nucleotides b_i, b_i represents A, C, G, or T. MR represents the *Mnt*-repressor; *S*-MR is the DNA-protein complex. Under the assumption that a protein can never be occupied by more than one DNA sequence at a time, the binding probability of a sequence S by *Mnt*-repressor molecules has a form of Fermi function [9],

$$P(S,\mu)=\frac{1}{1+\exp\left(\dfrac{-E_S-\mu}{k_BT}\right)}, \tag{1}$$

*Author whom correspondence should be addressed. Email address: qi@pku.edu.cn

TABLE I. Experimental data for a_i (defined as $\varepsilon_{b_i^*}-\varepsilon_{b_i}$) of nucleotide A,C,G,T in the position of i of sequence $S=b_1b_2\cdots b_L$ [11]. The best binding sequence is symmetric about position $i=9$. Because of this and that the *Mnt*-repressor is a tetramer of identical monomers, the data in the table is also symmetric (half data shown). For example, $a_1=0.27$ for nucleotide A at position $i=1$, the nucleotide T at $i=17$ also has the value $a_{17}=0.27$.

$b\backslash i$	1	2	3	4	5	6	7	8	9
A	0.27	0.76	2.36	0.67	2.36	2.36	0.0	3.2	0.74
C	1.3	1.1	3.37	1.21	0.0	0.0	1.67	0.0	0.0
G	0.0	0.0	0.0	0.32	2.19	4.38	2.02	2.53	0.0
T	1.2	0.81	2.19	0.0	1.06	1.67	2.53	1.85	0.74

where k_B is Boltzmann constant, μ is the chemical potential, and E_S is the binding energy of S to *Mnt*-repressor. Therefore $-E_S$ is the free energy of the complex S-MR (we assume that the free energy of free S be 0). The chemical potential is a selection threshold. An increase in μ lowers the threshold so that sequences with small binding energies are also likely to be selected. Practically, the chemical potential μ depends on the environment of DNA-protein binding. It can be controlled by the number of available proteins and the choice of the binding buffer.

We assume that E_S consists of two parts, i.e., specific binding energy ε_s and nonspecific binding energy ε_0:

$$E_S=\varepsilon_s+\varepsilon_0. \tag{2}$$

ε_s is determined by the binding details of the sequence S to *Mnt*-repressor, and therefore depends of specific nucleotides b_i in S. ε_0 is independent of specific sequences, which represents the contribution of the Coulomb interaction to the DNA-protein-binding affinity. It is identical for any possible sequence of L bases, and is a constant in our model. As it has been previously proved to be a good approximation for the *Mnt*-repressor system [12], we further assume that each nucleotide in the sequence contributes to the specific binding energy ε_s independently, i.e., $\varepsilon_s=\Sigma_{i=1}^{L}\varepsilon_{b_i}$, where ε_{b_i} is the energy contribution of the nucleotide b_i in the S sequence. Practically, the binding energy ε_s cannot be determined directly in experiment. However, if we appoint arbitrarily a sequence of nucleotides $S^*=b_1^*b_2^*\cdots b_L^*$ as a reference, then the discrepancy in binding energy of S^* from any sequence S, that is, $\varepsilon_{s*}-\varepsilon_s\equiv a_s$ with $a_s\equiv\Sigma_{i=1}^{L}a_i$ and $a_i\equiv\varepsilon_{b_i^*}-\varepsilon_{b_i}$, can be measured experimentally by the approach of point mutation [11]. An equivalent form of the specific binding energy ε_s can then be written as $\varepsilon_s=\varepsilon_{s*}-a_s=\Sigma_{i=1}^{L}(\varepsilon_{b_i^*}-a_i)$. In our model, we choose the DNA sequence that has the highest binding energy to *Mnt*-repressor as the reference. It reads S^* =GGGTCCACGGTGGACCC. This sequence is symmetrical about the central base $b_9^*=G$ in the sense that base G matches base C and base A matches base T. Actually, reference sequence S^* is the target of the *in vitro* evolution of *Mnt*-repressor. For this reference, the experimentally measured energy discrepancy a_i for all possible base matches of b_i^* with s_i are listed in Table I [11]. For all possible types of sequences with L bases, we calculated the landscape, i.e., the distribution $\Omega(a_s)$ of a_s, as depicted by squares in Fig. 1. It is obtained by calculating the number of DNAs that have their a_s fall in the range $[a_s,a_s+\Delta a_s]$ in the total 4^L variations of sequences.

Let $a_0=\varepsilon_{s*}-\varepsilon_0$. In our model, we make an approximation for the total binding energy by ignoring the smaller part of ε_0 or ε_s in E_S. E_S thus has the following form:

$$E_S=\begin{cases}\varepsilon_s & \text{if} \quad \varepsilon_s>\varepsilon_0\\ \varepsilon_0 & \text{if} \quad \varepsilon_s\leqslant\varepsilon_0.\end{cases} \tag{3}$$

This assumption is supported by previous experimental findings. Careful experiments with *lac*-repressor showed that the sequence-specific binding energy E_S can be replaced by the nonspecific energy ε_0 if the DNA sequence is sufficiently far from the best S^* sequence [13–15]. By taking Eq. (3), we have taken into account the effect of nonspecific binding affinity, which has been ignored in the model of Ref. [9]. In the selection probability, Eq. (1), we have three parameters: ε_{s*}, a_0, and μ. Denote $\mu_{eff}=\varepsilon_{s*}+\mu$, two independent parameters a_0 and μ_{eff} are left in the selection probability, which has the following form:

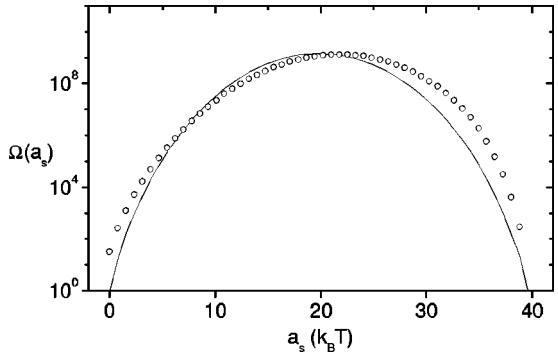

FIG. 1. The histogram of energy landscape $\Omega(a_s)$ calculated from Table I (the squares) and its coarse grain $\omega(m)$ calculated from Eq. (11) (solid line). $\Omega(a_s)$ and $\omega(m)$ represent the number of possible types of sequences whose a_s (refer to the text for its definition) fall in the range $[a_s, a_s+\Delta a_s]$. Δa_s for the histograms is $0.78k_BT$.

$$P(S,\mu)=\begin{cases}\dfrac{1}{1+\exp\left(\dfrac{a_S-\mu_{eff}}{k_BT}\right)} & \text{if } a_S<a_0\\[2ex] \dfrac{1}{1+\exp\left(\dfrac{a_0-\mu_{eff}}{k_BT}\right)} & \text{if } a_S\geqslant a_0.\end{cases} \quad (4)$$

As the Appendix shows, $\mu_{eff}=k_BT\ln(K_{S*}c_{Mf})$, where K_{S*} is the binding constant of the best sequence S^* and c_{Mf} is the concentration of free *Mnt*-repressor in the equilibrium state. K_s^* is affected by the reaction temperature T and pH value, and has a magnitude typical of 10^{11} M^{-1} [16]. c_{Mf} is determined by $c_{Mf}+\Sigma_S P(S,\mu)c_{St}=c_{Mt}$, where c_{St} and c_{Mt} are, respectively, the concentration of total DNA and *Mnt*-repressor. The dependence of c_{Mf} on c_{St},c_{Mt}, and K_s^* is in fact complicated and is not the interest of this study. We simply assume $c_{Mf}\approx c_{Mt}$, which is achieved when $c_{Mt}\gg\Sigma_S c_{St}$. Suppose the reactive buffer and the number of total *Mnt*-repressor remain unchanged during the experiment, μ_{eff} is a constant in the evolution process. Practically, *Mnt*-repressor concentration can vary from 10^{-11} M to 10^{-7} M [11,16], with $K_S^*\approx 10^{11}$ M^{-1} [16], the corresponding μ_{eff} ranges from 0 to $10k_BT$.

Using the above model, we carried out stochastic simulations of the *in vitro* evolution. Except otherwise stated, the model parameters were chosen as the following: $N=10^6$, $L=17$, $\nu_0=10^{-4}$, and $I=10$. The temperature T was fixed at 300 K which was typical in real experiment. In the simulation, we first generated 10^6 random DNA sequences of 17 nucleotides, with each base having an equal probability to be A,C,G, or T. The DNA population was then amplified 1000 times (10 cycles of ''PCR''). At each run of duplication, we generated a random number which is uniformly distributed in [0,1] for each nucleotide in all DNA. If the number is less than ν_0, then a mutation takes place on this nucleotide; the nucleotide was altered to other three different nucleotides with equal probability. Otherwise, the nucleotide was kept unchanged. After 10 cycles of PCR, a selection was made with each available sequence being selected according to the selection probability, Eq. (4). In order to recover the initial population, 10^6 sequences were random sampled from the selected pool. With a proper washing condition, this process can be realized in experiments by using a part of selected samples instead of all selceted samples. We investigated the dynamics of *in vitro* evolution by carrying out the above duplication-and-selection processes numerically.

III. SIMULATION

We scan the parameters a_0 in $[4k_BT,40k_BT]$ and μ_{eff} in $[0,10k_BT]$. Starting from a population of random sequences, the evolution always has a single destination: It always converges to the state in which most DNAs in the population are of S^*-type sequence.

To characterize the dynamics of the DNA population during evolution, we need to define a few parameters. First, the distance of a sequence $S=b_1b_2\cdots b_{17}$ to the final S^*

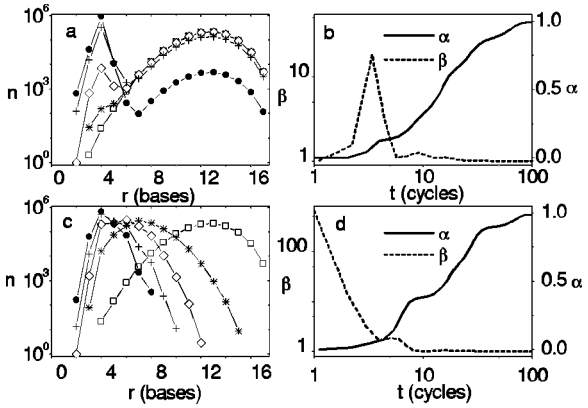

FIG. 2. The time evolution of distance distributions $n(r)$ for typical simulation runs that exhibit the phenomenon of crossover (a) and noncrossover (c). The symbols □, *, ◇, +, and ● represent the time $t=0$, 1, 2, 3, and 4, respectively. In (a), the distribution $n(r)$ at $t=3$ makes a jump to the state of $t=4$; while in the noncrossover case (c), $n(r)$ evolves smoothly. (b) and (d) depict the time dependence of the mean selection probability α and its growing rate β for the runs of (a) and (c), respectively. The crossover is best manifested by the peak of the curve for β. Other control parameters are: $\mu_{eff}=0$, $a_0=7k_BT$ for (a) and (b); $\mu_{eff}=0$, $a_0=40k_BT$ for (c) and (d).

=GGGTCCACGGTGGACCC sequence. We count the number of nucleotides in S which are different from S^* and denote it with r. The state of the population at the end of t cycles of duplication-mutation-selection can be roughly characterized by the distance distribution $n(r)$ as a function of r in the population. $n(r)$ describes how many sequences have the distance r to the S^* sequence in the population. Second, we denote the fraction that S-type DNA occupies in the population at time t with $f(S,t)$. Also, for a better characterization of the population during evolution, we introduce $\alpha(t)$ and $\beta(t)$:

$$\alpha(t)\equiv\sum_S P(S,\mu)f(S,t),$$

$$\beta(t)\equiv\frac{\alpha(t)}{\alpha(t-1)}. \quad (5)$$

$\alpha(t)$ is the averaged binding probability of the population, while $\beta(t)$ is a measure of the changing rate of the average binding ability, because $\beta(t)-1$ is exactly the changing rate of $\alpha(t)$.

By adjusting the parameters of μ_{eff} and a_0, we observe two types of dynamics: crossover and noncorssover. Typical crossover and non-crossover processes are shown in Figs. 2(a) and 2(c), respectively. They show the distance distribution $n(r)$ of the population at different evolution time t. One notes that in a crossover process [Fig. 2(a)], there is a sudden change of the distribution $n(r)$ during the evolution, revealing a sudden fast decrease in the average distance in the population. On the other hand, in a noncrossover case, the distance distribution $n(r)$ evolves steadily. Our calculation shows that with μ_{eff} in the range of 0 and $10k_BT$, if a_0

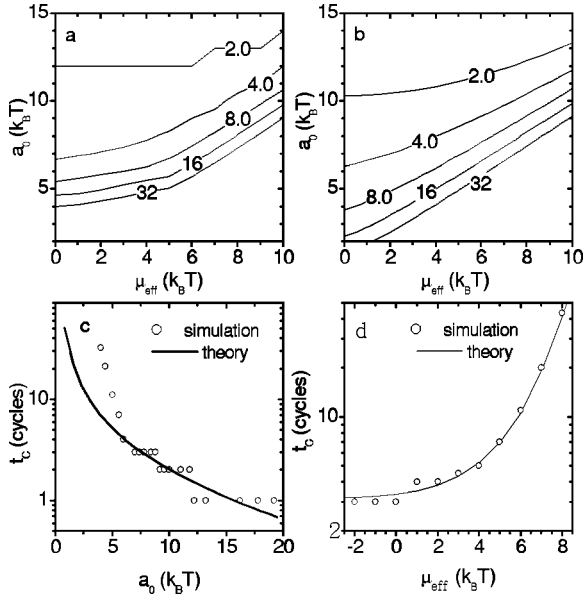

FIG. 3. The contour lines of characteristic crossover time t_c in the space of control parameters $\mu_{eff}-a_0$ obtained from simulations (a) and theory (b). (c) and (d) show the dependence of t_c on a_0 (with $\mu_{eff}=0$) and μ_{eff} (with $a_0=7k_BT$), respectively. Circles represent simulations and solid lines represent theory. The theoretical result is produced by calculating $\partial\beta(t)/\partial t|_{t_c}=0$.

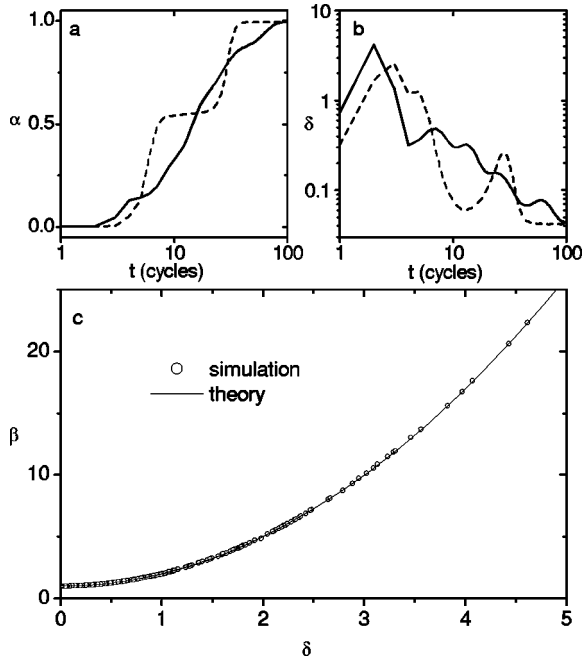

FIG. 4. Typical examples of two qualitatively different simulation runs found in stochastic simulations [the solid and dashed lines in (a) and (b)]. The dependence of β on the standard deviation δ of the selection probability α that follows Fisher's theorem of natural selection (c). The simulation result (squares) agrees very well with the function $\beta=1+\delta^2$ (solid line). Parameters: $\mu_{eff}=0$ and $a_0=7k_BT$.

$\leqslant 12k_BT$, a crossover takes place; if $a_0 \geqslant 14k_BT$, smooth evolution (noncrossover) occurs; and if $a_0 \in [12k_BT, 14k_BT]$, the occurrence of the crossover strongly depends on the chemical potential μ_{eff}: the crossover takes place only if μ_{eff} is larger than a certain lower limit value that is determined by the value of a_0. Figure 2(b) shows the corresponding time dependences of $\alpha(t)$ and $\beta(t)$ in a crossover case. At $t=0$, the initial random DNA population has an average distance of 13.0. During $t=1$ and 2, $\beta(t)$ grows very fast and the fraction of the sequence having four different bases to S^* sequence is found to grow abruptly in the population. However, these sequences are still in a minority (less than 1% in the population). At $t=3$, they become quite considerable in number (around 30% in the population). In the meantime, the average distance drops to 10.0. After this event, $\beta(t)$ undergoes a sudden drop while $\alpha(t)$ still grows steadily. This crossover process, best demonstrated by the peak of the $t\sim\beta(t)$ curve, can be found in a large range of parameters. The result of the simulation is consistent with the real DNA *in vitro* evolution experiments of Dubertret *et al.* [10]. Their experiment revealed an abrupt decrease in the average distance at the end of 5 cycles of PCR-mutation selection, and the average distance drops suddenly from 9 to 3. In contrast with Figs. 2(a) and 2(b), Figs. 2(c) and 2(d) show a noncrossover situation when $a_0=40k_BT$. In this case, $\beta(t)$ decreases monotonously without any abrupt behavior. Since sequence S^* has the largest selection probability $P(S^*,\mu)$, the final evolution result is still a population of S^* sequence, so that $\alpha(t)/P(S^*,\mu)$ always approaches to 1.0 as time proceeds.

The characteristic time of crossover t_c, defined as the time when the peak of $t\sim\beta(t)$ curve appears, is found to be affected by both parameters of μ_{eff} and a_0. Usually, for certain values of μ_{eff} and a_0, different simulation runs have different values of t_c. We thus calculate the mean ($\overline{t_c}$) of t_c on a number of simulation runs. The dependence of $\overline{t_c}$ on μ_{eff} and a_0 is summarized in Fig. 3(a), where the contour lines of $\overline{t_c}$ in μ_{eff}-a_0 space are plotted. The contour line of $\overline{t_c}=2$ distinguishes crossover and noncrossover evolution processes. Crossover appears if $\overline{t_c}\geqslant 2$; it becomes noncrossover if $\overline{t_c}=1$. The figure indicates that $\overline{t_c}$ is a decreasing function of a_0 and an increasing function of μ_{eff}. Figures 3(c) (circles) and 3(d) (circles) clearly demonstrate such two effects.

From different simulation runs carried out with various values of μ_{eff} and a_0, we recognized two qualitatively different types of trajectories. As shown in Fig. 4(a), the evolution path of the solid line is a of gradual changing trajectory, typically found in our simulations. It indicates that the average selection probability $\alpha(t)$ in the population is a steady growing function of time. The other type of typical run is represented by the dashed line in Fig. 4(a). It shows that the evolution process is divided into three slowly changing planar stages alternated with two fast growing stages. A fast growth in $\alpha(t)$ is typically preceded by a stagnant process. We calculated the time dependence of $\delta(t)$, i.e., the relative standard deviation of $P(S,\mu)$ at time t in the population. Figure 4(b) shows the curves of $\delta(t)$ for the two typical runs

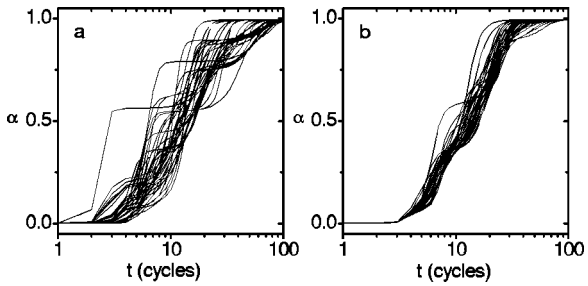

FIG. 5. Forty evolution trajectories selected arbitrarily from an ensemble of 200 simulation runs. They evolve from 40 different random initial DNA populations (a) and from an identical initial population (b), respectively. Parameters are the same as in Fig. 4.

in Fig. 4(a). Both are drastically undulant curves that exhibit frequent convergence and divergence in the population. This is the typical dynamical pattern we find in the *in vitro* evolution processes.

Biologically speaking, the selection probability $P(S,\mu)$ is a measure of fitness of an organism to the environment, and α represents the mean fitness of the population. The standard deviation δ of α in the population is actually an indicator of the diversity in the population. As early as over 90 years ago, Fisher discovered a fundamental theorem of natural selection in evolutionary biology: *The rate of increase in fitness of any organism at any time is equal to its genetic variance in fitness at any time* [17]. This biological law has been recovered in our numerical simulations of *in vitro* evolution. Figure 4(c) shows that the changing rate of α at any instance of time is just the square of the diversity δ or the variance of the fitness. The open squares in the figure are results of simulation and the solid line represents the function $\beta=1+\delta^2$. They agree very well.

We now fix the values $\mu_{eff}=0$ and $a_0=7k_BT$ and investigate the statistical properties of an ensemble of simulation runs. Two-hundred different realizations of simulations are conducted. For each simulation run, the initial population is prepared by randomly selecting N sequences from the pool of 4^L DNAs. Figure 5(a) shows the evolution trajectories of the ensemble in terms of $\alpha(t)$. The paths start from a small regime that lies a little above zero in α and terminate with a convergence to $\alpha=1$. During their evolution processes, the trajectories distribute, however, very diversely. There does not exist any typical evolution passage where trajectories keep close together during evolution. To complete the journey from an initial condition, a quick run requires only about ten duplication-mutation-selection cycles in order to have 90% of the DNAs in the population to be S^*-type sequence, while a slow run needs about 100 cycles to cover the journey. The dispersion $D(t)$ or standard deviation of $\alpha(t)$ in the ensemble at a few sampled times are listed in Table II (D_1); $D(t)$ has a small value of 0.00016 at $t=0$, indicating that the trajectories keep close initially. At $t=2$, $D(t)$ grows to 2.6 and the paths have been randomly dispersed. From $t=5$ to 100, $D(t)$ decreases gradually to a very small value and the system gradually converges to the optimal S^* sequence. For the purpose of comparison, the ensemble of evolution runs

TABLE II. The standard deviation D of the mean selection probability α in the ensemble at a few sampled times. D_1 denotes the case when the ensemble evolves from different random initial DNA populations and D_2 is the situation where the ensemble evolves from an identical initial population. Parameters are the same with Fig. 5.

$D\backslash t$	0	2	5	10	20	100
D_1	0.000 16	2.6	0.83	0.35	0.16	0.016
D_2	0.0	0.0096	0.28	0.13	0.10	0.0002

were also simulated from a completely identical initial random DNA population. Figure 5(b) gives the results. Trajectories in the figure are still dispersed, but the extent is much less drastic than that of Fig. 5(a). Again no typical passage exists. From Figs. 5(a) and 5(b), we deduce that the dispersion degree in the evolution trajectories comes from the diversity of initial conditions and the randomicity of DNA mutations. The dispersion of $\alpha(t)$ for this case is also listed in Table II (D_2).

We finally check the effects of system size N and the mutation rate ν_0. To characterize a simulation run, the evolution time t_e defined as the number of evolution cycles needed for 90% DNAs in the population to become the target S^* sequence, is calculated. At a parameter configuration, the mean of t_e, i.e., $\overline{t_e}$, is calculated for a number of simulation runs. Figure 6(a) shows the effects of N on $\overline{t_e}$. One observes that the size effect is significant when N is small. There is a characteristic size N_c, below which t_e blows up. The reason is that when the system is too small, the probability for sequences with $a_s<a_0$ to appear in an initial population is slim, and the system has to wait a long time for these sequences to be produced through a small mutation rate. The $N\sim\overline{t_e}$ dependence has a long tail and $\overline{t_e}$ converges at the limit of large system size. The effect of mutation rate on $\overline{t_e}$ is depicted in Fig. 6(b). At small values of ν_0, the system usually needs over several hundred or even higher numbers of duplication-mutation-selection cycles in order to attain to the S^*-sequence-dominant state. However, the effect of the mutation rate ν_0 is not so significant when $\nu_0>10^{-7}$ in the range of parameters that we checked.

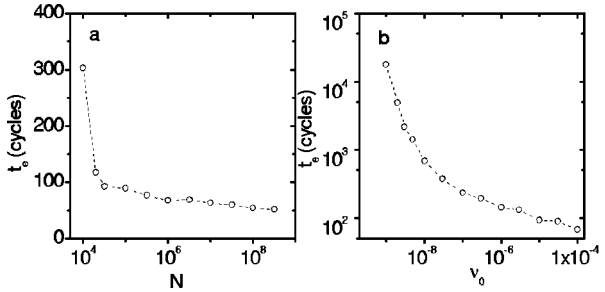

FIG. 6. Effects of the system size (a) and the mutation rate ν_0 (b) on the average evolution time t_e. Parameters: $\nu_0=10^{-4}$ for (a) and $N=10^6$ for (b). Other parameters are the same as in Fig. 4.

IV. ANALYSIS

We now give a theoretical analysis of the evolution dynamics of our model. In fact, $N=10^6$ means that the best sequence S^* has a small probability of 10^{-4} to appear in starting pool. For simplicity, we argue that $N=10^6$ is sufficiently large to contain sequences with different energy level, and small mutation rate ν_0 has insignificant influence on the population distribution of the DNA pool in the first several cycles. Thus we consider the limit $\nu_0=0$ and $N=\infty$. We denote the summation $[g]_t=\Sigma_S g(S)f(S,t)$, where $g(S)$ is an arbitrary function of S sequence and the sum is carried out over all possible types of sequences in a population [refer to the text that precedes Eq. (5) for $f(S,t)$]. Since there are only selections and no mutation takes place in the evolution, the sequence distribution function $f(S,t+1)$ at time $t+1$ can be calculated from the distribution $f(S,t)$ at t:

$$f(S,t+1)=\frac{P(S,\mu)}{[P]_t}f(S,t). \tag{6}$$

From the recurrence formula, $f(S,t)$ can be expressed in the initial distribution $f(S,0)$,

$$f(S,t)=\frac{[P(S,\mu)]^t}{[P^t]_0}f(S,0). \tag{7}$$

The forms of $\alpha(t)$ and $\beta(t)$ defined in Eq. (5) can be readily obtained from Eq. (7),

$$\alpha(t)=\frac{[P^{t+1}]_0}{[P^t]_0}, \tag{8}$$

$$\beta(t)=\frac{\alpha(t)}{\alpha(t-1)}=\frac{[P^{t+1}]_0[P^{t-1}]_0}{([P^t]_0)^2}. \tag{9}$$

Initially the fraction that any type of sequence S occupies in a population is uniform since the system size is infinite, and we have $f(S,0)=1/4^L$. From the selection probability, Eq. (4), we have

$$[P^t]_0=\frac{1}{4^L}\left(\sum_{a_s<a_0}\frac{\Omega(a_s)\triangle a_s}{\left[1+\exp\left(\frac{a_s}{k_BT}\right)\right]^t}+\sum_{a_s\geqslant a_0}\frac{\Omega(a_s)\triangle a_s}{\left[1+\exp\left(\frac{a_0}{k_BT}\right)\right]^t}\right). \tag{10}$$

Apparently, the mean $[P^t]_0$ are determined by the landscape $\Omega(a_s)$ and parameters a_0. $\Omega(a_s)$ is actually the energy density distribution for the initial population. The sum in Eq. (10) consists of two parts: $a<a_0$ and $a\geqslant a_0$.

In order to obtain an analytical form for $\beta(t)$, the experimental landscape shown in Fig. 1 should be approximated analytically. We make a coarse-grain for the experimental data in Table I. We divide all the types of free energies a_i ($a_i=\varepsilon_{b_i^*}-\varepsilon_{b_i}$) contributed possibly by a nucleotide in a sequence S uniformly into four classes. a_i can only take values of 0, ϵ, 2ϵ, or 3ϵ. In accordance with data in Table I, we take $\epsilon=0.78k_BT$. For a sequence S that has l_0 nucleotides with $a_i=0$, l_1 nucleotides with $a_i=\epsilon$, l_2 nucleotides with $a_i=2\epsilon$, and l_3 nucleotides with $a_i=3\epsilon$, while $l_0+l_1+l_2+l_3=L$. The sequence S thus has the free energy $a_s\equiv m\epsilon=(l_1+2l_2+3l_3)\epsilon$. The number of all possible types of sequences that satisfy $a_s=m\epsilon$ can be calculated to be

$$\omega(m)=\sum_{l_3=0}^{[m/3]}\sum_{l_2=\max(0,m-L-2l_3)}^{\min\{L-l_3,[(m-3l_3)/2]\}}C_L^{l_3}C_{L-l_3}^{l_2}C_{L-l_3-l_2}^{m-3l_3-2l_2}, \tag{11}$$

where the bracket [ ] is the operation of taking the integer part of a real number. The landscape $\omega(m)$ is the approximation of $\Omega(a_s)$. As depicted in Fig. 1, the solid curve for $\omega(m)$ agrees well with the squares of experimental data for $\Omega(a_s)$.

By virtue of Eqs. (10) and (11), $\beta(t)$ can be finally computed. By adjusting the parameters μ_{eff} and a_0, we find that the crossover phenomenon observed in our simulations shows up with small values of a_0. Figure 7(a) demonstrates such an example. The solid curve is the theoretical result, and the dashed line is calculated by averaging β on a bound of simulation runs. The characteristic crossover time (i.e., t_c, the time the peak locates) predicted by theory also agrees well with the simulations. Figure 7(b) depicts the case of the situation without crossover when a_0 takes large values ($a_0>12k_BT$). The crossover time t_c can be also determined by $\partial\beta(t)/\partial t|_{t_c}=0$. On the basis of Eqs. (9)–(11), the dependence of t_c on μ_{eff} and a_0 was calculated, as shown in Fig. 3(b). The contour lines of t_c qualitatively agree with simulations [Fig. 3(a)]. The solid lines of theoretical prediction in Figs. 3(c) and 3(d) are consistent with numerical simulations.

Fisher's theorem of natural selection can be readily derived with $\nu_0=0$. We consider that the population is large enough. At time t, the fraction of sequence S in the population is proportional to $P(S,\mu)f(S,t-1)$, that is,

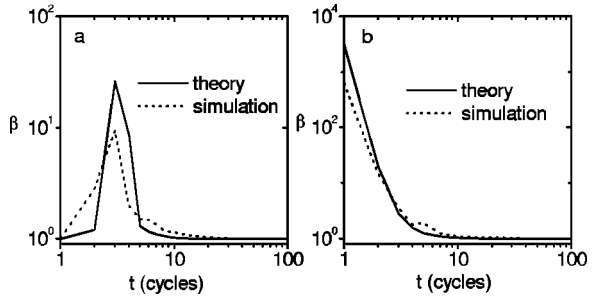

FIG. 7. The $\beta\sim t$ correspondence from simulation (dashed) and theory (solid). Theoretical result in (a) also manifests the crossover phenomenon found in simulations. The results of simulation are calculated by taking an average over 200 runs. Parameters: $\mu_{eff}=0$, $a_0=7k_BT$ for (a); $\mu_{eff}=0$, $a_0=40k_BT$ for (b). Other parameters are the same as in Fig. 4.

$$f(S,t)=\kappa P(S,\mu)f(S,t-1), \tag{12}$$

where κ is the coefficient. We integrate both sides of the above equation with respect to S and arrive at

$$\kappa\sum_S P(S,\mu)f(S,t-1)=\kappa\alpha(t-1)=\sum_S f(S,t)=1. \tag{13}$$

We have $\kappa=1/\alpha(t-1)$ and get the expression for $f(S,t)$,

$$f(S,t)=\frac{P(S,\mu)f(S,t-1)}{\alpha(t-1)}. \tag{14}$$

Therefore the mean binding probability takes the following form:

$$\alpha(t)=\sum_S P(S,\mu)f(S,t)=\frac{\sum_S P(S,\mu)^2 f(S,t-1)}{\alpha(t-1)}. \tag{15}$$

From the expression of $\alpha(t)$, it is easy to prove that

$$\beta(t)=1+\frac{\sum_S [P(S,\mu)-\alpha(t-1)]^2 f(S,t-1)}{\alpha(t-1)^2}$$
$$=1+\delta^2(t-1). \tag{16}$$

δ^2 is the variance of $P(S,\mu)$ and $\beta-1=\delta^2(t-1)$ is just Fisher's theorem of natural selection.

V. DISCUSSION

We have investigated the dynamics of competitive DNA *in vitro* evolution with *Mnt*-repressor numerically and analytically. The selection strength of the chemical potential μ_{eff} is regarded as an invariant during the evolution process, which is a practical case when the protein molecules are excessively present. Based on the experimental data for the sequence-specific binding energy and the corresponding landscape, the evolution process were simulated. By changing the selection strength μ_{eff} within the practical range of $[0,10k_BT]$, we demonstrated that the crossover process observed in the experiment [10] can take place only when a_0 is smaller than $14k_BT$. In fact, a small value of a_0 represents a case where nonspecific binding energy ϵ_0 is dominant. With large a_0, the contribution of nonspecific energy can be ignored, only specific energy takes effect. These results reveal that the nonspecific energy is responsible for the crossover phenomenon. Qualitatively speaking, the bigger the portion of initial sequences with nonspecific binding energy, the easier the occurrence of crossover. It suggests that in real experiments, the energy discrepancy of nonspecific energy ε_0 to the best specific energy ε_S^* might be at most $14k_BT$. We compare this magnitude evaluation with previous prediction [18]. In Ref. [18], Gerland *et al.* proposed that $E_{S*}-\varepsilon_0\approx k_BT\ln(\Gamma)$, where Γ is the size of genome that contains the DNA sequence. This relation was proved to be correct for all cases where nonspecific energy has been measured experimentally. In this light, $E_{S*}-\varepsilon_0\approx 16k_BT$ for *Mnt* because the *Samonella* genome size is $\approx 5\times 10^6$. Together with $E_{S*}\approx 25k_BT$ for $K_{S*}\sim 10^{11}M^{-1}$ [16], one can calculate that $a_0=\varepsilon_{S*}-\varepsilon_0=E_{S*}-2\varepsilon_0=2(E_{S*}-\varepsilon_0)-E_{S*}\approx 7k_BT$, which quantitatively agrees with our prediction. When a_0 is properly set, we showed that the crossover can take place in the first few evolution cycles, as was observed by experiment [10], with a strong selection force that can be quantitatively controlled by protein concentration.

With an ensemble of simulation runs, it was revealed that the evolution trajectories are drastically dispersed and there do not exist typical evolution passages where the trajectories keep close together. We thus speculate that diversification must be the key of the dynamics of evolution. With a coarse-grained simplification for the experimental data of specific binding energy, we obtained a simplified energy landscape for the system. Based on this simplification and the assumption that small mutations do not have a major effect on the population distribution of the DNA pool, Fisher's theorem of natural selection, which states that the growth rate of fitness of an organism is exactly its variance of fitness, is verified and put in an analytical expression [Eq. (16)]. This analytical formula is quantitatively consistent with the result of the computer simulation.

ACKNOWLEDGMENTS

We thank Dr. Fangting Li for his helpful advice. This work was supported by the ''863'' program of Department of Science and Technology, China.

APPENDIX

When the reaction $S+\mathrm{MR}\rightleftharpoons S-\mathrm{MR}$ attains the equilibrium state, we have $K_S=c_{Sb}/(c_{Sf}c_{Mf})$, where K_S is the binding constant of S, c_{Sb}, c_{Sf}, and c_{Mf} are, respectively, the concentration of binding SS, free S, and free *Mnt*-repressor. With $c_{Sb}/c_{Sf}=K_Sc_{Mf}$, the selection probability has the form

$$P(S,\mu)=\frac{c_{Sb}}{c_{Sf}+c_{Sb}}=\frac{1}{1+(K_Sc_{Mf})^{-1}}. \tag{A1}$$

Using the relation $K_S\propto\exp(E_S/k_BT)$, we have

$$K_S=K_{S*}\exp\left(\frac{E_S-E_{S*}}{k_BT}\right)=\begin{cases} K_{S*}\exp\left(\dfrac{-a_S}{k_BT}\right) & \text{if } a_S<a_0 \\ K_{S*}\exp\left(\dfrac{-a_0}{k_BT}\right) & \text{if } a_S\geqslant a_0. \end{cases} \tag{A2}$$

So that the final form of $P(S,\mu)$ is

$$P(S,\mu)=\begin{cases}\dfrac{1}{1+(K_{S*}\cdot c_{Mf})^{-1}\exp\left(\dfrac{a_S}{k_BT}\right)}=\dfrac{1}{1+\exp\left(\dfrac{a_S-k_BT\ln(K_{S*}\cdot c_{Mf})}{k_BT}\right)} & \text{if } a_S<a_0\\ \dfrac{1}{1+(K_{S*}\cdot c_{Mf})^{-1}\exp\left(\dfrac{a_0}{k_BT}\right)}=\dfrac{1}{1+\exp\left(\dfrac{a_0-k_BT\ln(K_{S*}\cdot c_{Mf})}{k_BT}\right)} & \text{if } a_S\geqslant a_0\end{cases} \tag{A3}$$

in which $k_BT\ln(K_{S*}c_{Mf})$ is exactly μ_{eff} of Eq. (4).

[1] D.S. Wilson and J.W. Szostak, Annu. Rev. Biochem. **68**, 611 (1999).
[2] Y.-Y. He, P.G. Stockley, and L. Gold, J. Mol. Biol. **255**, 55 (1996).
[3] N.M. Low, P. Holliger, and G. Winter, J. Mol. Biol. **260**, 359 (1996).
[4] C. Tuerk and L. Gold, Science **249**, 505 (1990).
[5] L.C. Bock *et al.*, Nature (London) **355**, 564 (1992).
[6] R. Pollock and R. Treisman, Nucleic Acids Res. **18**, 6197 (1990).
[7] D. Irvine, C. Tuerk, and L. Gold, J. Mol. Biol. **222**, 739 (1991).
[8] B. Vant-Hull, A. Payano-Baez, R.H. Davis, and L. Gold, J. Mol. Biol. **278**, 579 (1998).
[9] W. Peng, U. Gerland, T. Hwa, and H. Levine, Phys. Rev. Lett. **90**, 088103 (2003).
[10] B. Dubertret, S. Liu, Q. Ouyang, and A. Libchaber, Phys. Rev. Lett. **86**, 6022 (2001).
[11] D.S. Fields, Y.-Y. He, A.Y. Al-Uzri, and G.D. Stormo, J. Mol. Biol. **271**, 178 (1997).
[12] G.D. Stormo, S. Strobl, M. Yoshioka, and J.S. Lee, J. Mol. Biol. **229**, 821 (1993).
[13] M.T. Record, Jr., P.L. deHaseth, and T.M. Lohman, Biochemistry **16**, 4791 (1977).
[14] R.B. Winter and P.H. von Hippel, Biochemistry **20**, 6948 (1981).
[15] D.E. Frank *et al.*, J. Mol. Biol. **267**, 1186 (1997).
[16] A.K. Vershon *et al.*, J. Mol. Biol. **195**, 311 (1987).
[17] R.A. Fisher, in *The Genetical Theory of Natural Selection* (Oxford University Press, Oxford, 1999), p. 35.
[18] U. Gerland, J.D. Moroz, and T. Hwa, Proc. Natl. Acad. Sci. U.S.A. **99**, 12 015 (2002).