# Simulation and Analysis of *in vitro* DNA Evolution

Morten Kloster[1,2] and Chao Tang[2,3]

[1]*Department of Physics, Princeton University, Princeton, New Jersey 08544, USA*
[2]*NEC Laboratories America, 4 Independence Way, Princeton, New Jersey 08540, USA*
[3]*Center for Theoretical Biology, Peking University, Beijing 100871, China*

(Received 12 February 2003; published 22 January 2004)

We study theoretically the *in vitro* evolution of a DNA sequence by binding to a transcription factor. Using a simple model of protein-DNA binding and available binding constants for the *Mnt* protein, we perform large-scale, realistic simulations of evolution starting from a single DNA sequence. Varying the evolution parameters reveals three different regimes characterized by distinct evolutionary behaviors, and for each regime we find analytical estimates which agree well with simulation results. We also study how the details of the DNA-protein interaction affect the evolution.

DOI: 10.1103/PhysRevLett.92.038101 PACS numbers: 87.10.+e, 87.23.Kg

The concept of evolution has not only fundamentally shaped our view of biology, but also found rich and profound applications in bioengineering and biotechnology. In particular, *in vitro* evolution has been widely used to evolve DNA [1], RNA [2], and proteins [3]. For the evolution of DNA via binding to a protein, the relation between the genotype (DNA sequence) and the phenotype (binding affinity) is direct and simple. If the binding constants are known for various DNA sequences, the selection process can be modeled quantitatively and it can then serve as a model system for quantitative analysis of molecular evolution. A recent Letter by Peng *et al.* [4] analyzes the evolution dynamics of a model with weak competition and high mutation rates, and shows an exponential approach to the equilibrium state. In this Letter we study the (very different) behavior of a model in the more experimentally relevant regime of strong competition and low mutation rates. We explore a large range of experimentally accessible parameters and find various regimes with very distinct evolution dynamics. A recent experiment by Dubertret *et al.* [5] shows how such evolution can be carried out (but does not test our model [6]).

In this Letter, we study theoretically the *in vitro* evolution of DNA sequences via binding to the *mnt* repressor protein. DNA-*mnt* is perhaps the best experimentally characterized system of sequence-specific DNA-protein binding [8–10]. It has been demonstrated that the binding energy of a sequence can be approximately decomposed as the sum of contributions from the individual bases, all of which have been estimated experimentally [8,10]. This *additive* form of binding energy greatly simplifies the analysis—it enables us to perform realistic large-scale simulations as well as to obtain analytic solutions and estimates in various cases. (The values of the contributions to the binding energy are not qualitatively important.)

We assume that the binding energy between a DNA molecule and the *mnt* protein is of the form $E(S) = \sum_i \epsilon_i(S_i)$, where $S_i \in A, C, G, T$ is the base at the $i$th position of the DNA sequence and $\epsilon_i(S_i)$ is the contribution to the binding energy from this position, for which we use the value in Ref. [8]. The relative binding constants are then $K(S) = \prod_i K_i(S_i) = \prod_i e^{-\beta\epsilon_i(S_i)}$ [11]. We start with a population size $N$ of identical DNA molecules of a starting sequence (SS) that is significantly different from the wild type (WT) [7], avoiding the potential problem of enrichment [6]. An iteration of the evolutionary process consists of an amplification with mutation followed by a selection. During amplification the population is doubled [e.g., by polymerase chain reaction (PCR)], with a (low) error rate of $r$ per base for the new copies. The population is then subject to a selection process via binding to the *mnt* protein. Each DNA molecule is selected with a probability $\frac{1}{1+e^{\beta[E(S)-\mu]}}$, where the chemical potential $\mu$ is chosen such that the expected number of selected molecules is $N$. [We limit our analysis here to a single duplication (cycle of PCR) per iteration, as this case is particularly tractable. Aspects of the general case are discussed elsewhere [12], as are consequences of various assumptions we have made.]

The binding site for the *mnt* repressor consists of 17 important base pairs, at positions 3 through 19. For our SS we chose, by random mutations (any choice yields similar behavior, as long as specific binding dominates), a sequence that differs from WT at $m = 6$ positions (Table I). We call a specific sequence of mutations that take a SS to the WT an evolution path. Each path contains

TABLE I. The starting sequence SS. $\Delta K = e^{-\beta\Delta E}$ is the ratio of the binding constants due to the specified mutation type (the given base substitution at the given position).

| | Changes from WT | | | | | |
|---|---|---|---|---|---|---|
| Position | 4 | 7 | 9 | 10 | 13 | 15 |
| WT base | G | C | A | C | T | G |
| SS base | T | A | C | T | C | C |
| $\Delta K$ | 2.13 | 10.0 | 5.0 | 5.5 | 7.2 | 8.3 |

  

the six required mutations in some order, and may contain additional mutations ($m! = 720$ "minimum paths" contain only the six required mutations). A simulation ends when at least 90% of the DNA molecules are WT, giving the number $t_M$ of iterations required and the number of WT molecules $n_{t_M}^{\pi}$ coming from each path $\pi$ [13].

Some of the key quantities from a single simulation run are the fraction of WT that was produced through minimum paths, $f_{\min}^{\mathrm{WT}}$, the number of minimum paths used, $n_{\min}$, and the fraction of the WT produced through the single best path in the simulation, $f_{\mathrm{best}}^{\mathrm{WT}}$. Figure 1 shows how these quantities depend on the population size (averaged over many simulations): $f_{\min}^{\mathrm{WT}}$ is small for very small $N$ and is close to 1 for large $N$, with a fairly sharp transition, whereas $f_{\mathrm{best}}^{\mathrm{WT}}$ slowly decreases from 1 with increasing $N$. This indicates that we may find qualitatively different behavior for small and large population sizes.

Let us first consider $N = 1$, i.e., at each iteration a single DNA molecule is duplicated and one of the two molecules is selected. If there are no mutations during amplification, nothing interesting can happen. If there is a mutation $i$, the chance of selecting the mutant is $\frac{\Delta K_i}{1+\Delta K_i}$, i.e., high for good mutations ($\Delta K_i > 1$) and low for bad mutations. The DNA molecule performs a biased random walk, making a step whenever a mutant is selected. Specifically, these transition probabilities describe a random walk in the energy landscape given by the binding energy—in equilibrium, $\mathrm{Prob}(S) \propto e^{-\beta E(S)}$.

Now consider a population size $1 < N \ll 1/r$. The chance of having a mutation in any single iteration is very low. When a mutation occurs, it will almost certainly either "die out" (disappear from the population) or spread through ("replace") the whole population before the next mutation occurs; most of the time the population consists of $N$ identical DNA molecules [14]. During selection, the chance of choosing a particular combination of DNA molecules is proportional to the product of their binding constants. We keep exactly half the DNA at each selection, and the probabilities $p_{\pm}^{N}(k)$ that $k$ identical mutants of type $i$ in a population of size $N$ will replace the population ($p_+$) or die out ($p_-$) can be shown to satisfy $p_+^N(N-k) = (\Delta K_i)^{2k} p_-^N(k)$. Including the probability for a mutant to be selected in the first place, and using $p_+^N(k) + p_-^N(k) = 1$, we find the probability $P_i$ for a single mutation of type $i$ to replace the population:

$$P_i = \frac{\Delta K_i}{\Delta K_i + 1} p_+^N(1) = \frac{(\Delta K_i)^{2N} - (\Delta K_i)^{2N-1}}{(\Delta K_i)^{2N} - 1} \approx \begin{cases} (1 - \frac{1}{\Delta K_i}), & \Delta K_i > 1, \\ 0, & \Delta K_i < 1, \end{cases} \quad (1)$$

where the approximation is valid for $|2N \log \Delta K_i| \gg 1$. The population again performs a random walk, but now the energy landscape is $2N - 1$ times the binding energy of a single DNA molecule—in equilibrium, $\mathrm{Prob}(N \text{ copies of } S) \propto e^{-\beta(2N-1)E(S)}$.

The average time needed to improve the DNA pool by one base relative to WT can now be estimated as

$$\langle T \rangle \approx \left[ \sum_i \frac{Nr}{3} \left( 1 - \frac{1}{\Delta K_i} \right) \right]^{-1} + \frac{\log(N)}{\log(\frac{2\Delta K}{1+\Delta K})}, \quad (2)$$

where $\Delta K$ is a typical value for $\Delta K_i$. The first term is the average time needed to create a "seed" mutation—we sum over all possible correct mutations the rate $\frac{Nr}{3}$ at which each mutation occurs times the chance that it survives—and the second term is the time required to replace the population ($\frac{2\Delta K_i}{1+\Delta K_i}$ is the effective amplification for mutant type $i$), which is negligible for small $N$. Since the first term, which dominates, is inversely proportional to the distance from WT (i.e., the number of terms in the sum), the average speed of the DNA pool is proportional to that distance. Figure 2(a) shows the distance from WT as a function of time, and except in the beginning, it can be almost perfectly fitted to an exponential—this is similar to the result in [4], which is for a very different regime. The corrections for the beginning are precisely what we would expect from the second term: It limits the speed and causes a short delay. Our result for the evolution speed in the random walk (RW) regime is

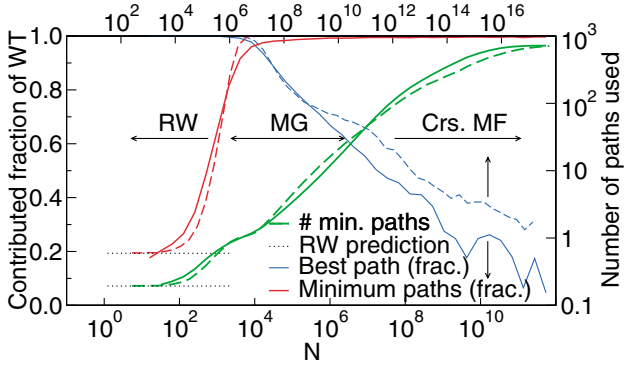


FIG. 1 (color). The average fraction of WT contributed by the best path, the total fraction from all minimum paths, and the number of different minimum paths used, for different population sizes $N$. $r = 10^{-4}$ for solid lines, and $r = 10^{-7}$ for dashed lines. The various regimes (random walk, middle ground, and crossover to mean field) are indicated.

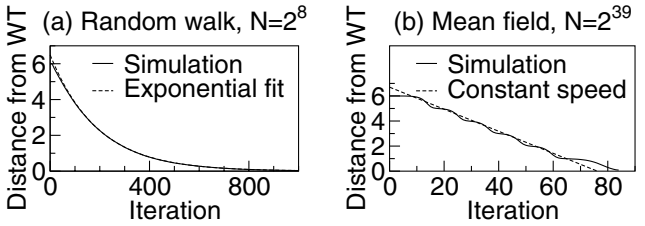


FIG. 2. Average number of different bases between DNA pool and WT as a function of time. $r = 10^{-4}$.

very similar to the one found in [14] for a birth-rate model: As long as the second term of Eq. (2) is negligible, the evolution speed of the DNA population is proportional to the mutation rate and to the population size $N$.

Given a sequence $S$, mutation $i$ will be the next surviving mutation with probability (for $Nr \ll 1$ and large $N$)

$$P_{RW\mathrm{mut}}(S, i) = \frac{1 - \frac{1}{\Delta K_i}}{\sum_j 1 - \frac{1}{\Delta K_j}}, \tag{3}$$

where $j$ runs over all possible good mutations to $S$. $P_{RW\mathrm{path}}(\pi) = \prod_n P_{RW\mathrm{mut}}(S_n^\pi, \pi_n)$ is then the chance of following a given path $\pi$. Figure 3(a) compares these predicted values with simulation results (which are Poisson distributed), and the agreement is excellent.

The approach to the small $Nr$ limit for the observed total fraction of minimum paths is shown in Fig. 1. The random walk approach is not accurate if the second term of Eq. (2) exceeds the first term: A mutation will not have time to replace the population before the next favorable mutation occurs.

For sufficiently large $N$, we expect mean field (MF) behavior—there is no significant difference between one experiment/simulation and another, and each path contributes a fixed fraction to the total WT DNA produced. In principle, in the limit of $N \to \infty$ the fraction of population $S$ at time $t$, $f(S, t)$ can be traced from iteration to iteration and the chemical potential determined from the selection criterion $\sum_S f(S, t + 1/2)/\{1 + \exp[E(S) - \mu(t)]\} = \frac{1}{2}$, where $f(S, t + 1/2)$ is the fraction of $S$ just after the amplification but before the selection. This would require tracing all possible sequences—a tedious and often impractical task. In practice, it suffices to restrict ourselves to a limited set of paths [15]. With a fixed set of paths, we can find the fraction of the DNA at each step of each path at each iteration, as well as the chemical potential used for each selection. The calculations proceed the same way as for the finite $N$ simulations, except there is no randomness involved, and we discard the DNA that departs from the chosen paths.

Let $n_0^{\mathrm{SS},\pi}$ be the number of WT molecules produced from a single SS molecule through any specific path $\pi$ in an experiment/simulation. Once we know all the chemical potentials $\mu(t)$ used for the selections, we can use a set of recursion relations to find both the average $\langle n_0^{\mathrm{SS},\pi}\rangle$, the variance $\mathrm{Var}(n_0^{\mathrm{SS},\pi})$, and the probability $P_+(n_0^{\mathrm{SS},\pi}) = P(n_0^{\mathrm{SS},\pi} > 0)$ that one SS molecule at time $t = 0$ will yield some nonzero amount of WT at $t_M$ through $\pi$ [12]. In the mean field regime, the amount of WT produced through each minimum path should be relatively constant from one experiment to the next, which gives us a lower bound for the population size:

$$N > \max_\pi \frac{\mathrm{Var}(n_0^{\mathrm{SS},\pi})}{\langle n_0^{\mathrm{SS},\pi}\rangle^2}. \tag{4}$$

These bounds vary from $1.1 \times 10^{19}$ ($r = 10^{-4}$) to $1.4 \times 10^{38}$ ($r = 10^{-7}$ and 15 cycles of PCR per iteration) [12]. The dependence on $r$ is roughly $N \sim r^{-m}$.

The speed of evolution in the mean field regime is easily estimated—half the population is removed during each selection, thus $\mu$ will be close to the median binding energy in the population, and any DNA with significantly higher binding energy than the majority will almost certainly survive. After the very first iteration, a "seed fraction" $r/3$ of the DNA will have each of the $m$ "correct" mutations. In the following iterations, almost all of this improved DNA will survive the selections, and the amount of improved DNA will thus roughly be doubled in each iteration. After

$$T_0 \approx 1 + \frac{\log(\frac{3}{mr})}{\log(2)} \tag{5}$$

iterations, the improved DNA will have replaced the original population; i.e., most of the DNA will only have $m - 1$ errors relative to WT.

Once the whole population has improved by one base, the process is repeated. However, there have already been $T_0$ iteration in which improved DNA could improve further through mutations, and this even better DNA was amplified at the same rate as the regular improved DNA, i.e., the seed fraction is now $\frac{1}{2}\frac{T_0(m-1)r}{3}$ ($\frac{1}{2}$ because PCR introduces mutations only in the copy). The DNA pool will thus improve by one base roughly every

$$T(m') \approx 1 + \frac{\log(\frac{3}{m'Tr})}{\log(2)} \tag{6}$$

iterations, where $m'$ is the number of errors left. Some of the improved DNA will be lost during selection, and the *effective* number of errors is much lower than the actual number $m'$—these and other corrections can be addressed by considering an infinite length model [16].

The resulting evolution speed is almost constant [Fig. 2(b)]; using $T(2)$ from Eq. (6) gives a very good fit. The first improvement takes somewhat longer, as expected, and so does the last improvement: The error at position 4 (usually the last error to be corrected) has a very low $\Delta K$, i.e., low effective amplification.

The argument used for the evolution speed in the MF regime is qualitatively valid for all $N > \frac{1}{r}$; thus we expect a smooth transition from the RW evolution behavior,

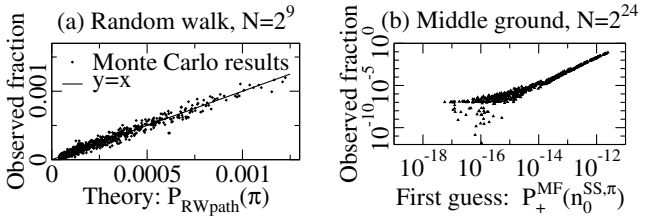


FIG. 3. Average contribution of individual minimum paths—observed vs predicted. $r = 10^{-7}$. (a) $2^{17}$ and (b) $2^{23}$ runs.

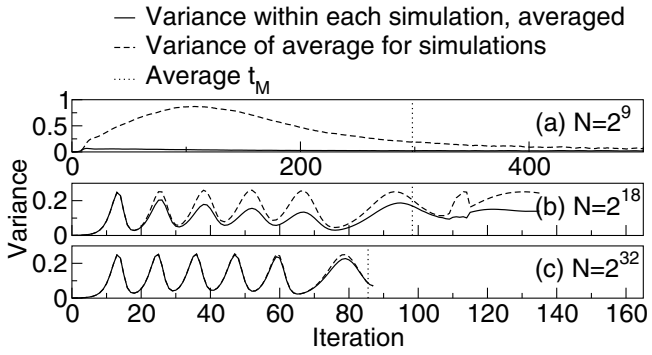


FIG. 4. The variance of the number of errors relative to WT as a function of time. $r = 10^{-4}$ for all graphs.

random with on average exponential approach, to the MF behavior, constant speed. However, there is a significant region where $f_{\min}^{\mathrm{WT}} \approx 1$ and $f_{\mathrm{best}}^{\mathrm{WT}} > 0.5$; i.e., usually a single minimum path dominates (Fig. 1). In this region, the average contribution of each path, $\langle n_{t_M}^{\pi} \rangle$, depends strongly on how often that path dominates. As the overall evolution behavior resembles that of the MF regime, we can use as a first guess the probabilities $P_{+}^{\mathrm{MF}}(n_0^{\mathrm{SS},\pi})$ from the large $N$ discussion (with the superscript MF to emphasize that these are mean field calculations). Figure 3(b) shows the results from simulations plotted against this estimate, and $\langle n_{t_M}^{\pi} \rangle$ is indeed closely related to $P_{+}^{\mathrm{MF}}(n_0^{\mathrm{SS},\pi})$ (the relationship is not linear), at least for the most probable paths—for the less probable paths, statistical errors are large. $\langle n_{t_M}^{\pi} \rangle$ varies far more here than in the RW and MF regimes, and we consider this region to be a third parameter regime, the middle ground.

The middle ground region corresponds fairly well to $\frac{1}{r} < N < [\sum_{\pi} P_{+}^{\mathrm{MF}}(n_0^{\mathrm{SS},\pi})]^{-1}$; i.e., the regime ends approximately when the population is large enough that, using the mean field chemical potentials, we would expect to find at least some WT in most simulations. There is a very large crossover region between the MF regime and the middle ground, and a smaller region between the middle ground and the RW regime (Fig. 1).

It is difficult to completely and directly test the above theoretical analysis experimentally—one would need to sequence a large number of DNA. A more practical choice is to consider only the distance from a sequence to WT, i.e., the number of positions at which they differ, and study its variance in different regimes. In the random walk regime, at most times the DNA pool consists of only a single sequence, thus the average variance for a population snapshot is almost zero, but the variance between runs can be very large. As we increase the population size, the variance within a run increases somewhat, but the variance between runs decreases drastically, and above the middle ground we have almost perfect coherence (Fig. 4)—in particular, there are specific times at which almost the whole population has each given distance from WT [Figs. 2(b) and 4(c)].

Our simulation and analysis show that in the simple case of additive binding energy the evolution behavior of DNA-protein binding can be understood quantitatively and rather completely. Depending on the population size and the mutation rate, the evolutionary process exhibits distinct behaviors in three parameter regimes, and for large populations the behavior is very different from those observed in other models [4,17]. Our results are fairly general as long as the potential is mainly additive and can be used to make sense of experimental data. The additivity of the binding energy gives rise to a smooth landscape, which greatly simplifies the analysis. The inclusion of a small perturbative nonadditive part to the potential would not change the picture, but would nonetheless provide insights to the cases of more general potentials and fitness functions.

We thank Qi Ouyang and Terry Hwa for very helpful discussions.

[1] J. A. Bittker, K. J. Phillips, and D. R. Liu, Curr. Opin. Chem. Biol. **6**, 367 (2002).
[2] L. F. Landweber, Trends Ecol. Evol. **14**, 353 (1999).
[3] *Evolutionary Protein Design*, edited by F. H. Arnold (Academic Press, San Diego, 2001).
[4] W. Peng, U. Gerland, T. Hwa, and H. Levine, Phys. Rev. Lett. **90**, 088103 (2003).
[5] B. Dubertret, S. Liu, Q. Ouyang, and A. Libchaber, Phys. Rev. Lett. **86**, 6022 (2001).
[6] Enrichment (i.e., sequences very close to WT [7] in the initial pool being amplified exponentially) was crucial for the observed evolution dynamics in [5], as the authors started from a mix of random DNA sequences.
[7] We let WT denote the highest affinity sequence, although this is not the actual wild type [5,8].
[8] D. S. Fields, Y. He, A. Al-Uzri, and G. D. Stormo, J. Mol. Biol. **271**, 178 (1997).
[9] G. D. Stormo and D. S. Fields, Trends Biochem. Sci. **23**, 109 (1998).
[10] T.-K. Man and G. D. Stormo, Nucleic Acids Res. **29**, 2471 (2001).
[11] DNA molecules may also bind to the protein in a nonspecific manner [M. T. Record, Jr., P. L. deHaseth, and T. M. Lohman, Biochemistry **16**, 4791 (1977)], $K_{tot}(S) = K(S) + K_{\mathrm{NSB}}$. This binding dominates for DNA sequences very far from WT, but is negligible for the case studied in this Letter [12](we include it in simulations).
[12] M. Kloster and C. Tang, cond-mat/0301372.
[13] For molecules simultaneously mutated at multiple positions, we randomly assign an order to the mutations.
[14] D. A. Kessler, H. Levine, D. Ridgway, and L. Tsimring, J. Stat. Phys. **87**, 519 (1997).
[15] For large $N$, almost all the WT comes from minimum paths (Fig. 1). To be safe, we allow one erroneous mutation and verify that this is a minor correction.
[16] M. Kloster (unpublished).
[17] D. Ridgway, H. Levine, and D. A. Kessler, J. Stat. Phys. **90**, 191 (1998).