Dynamics of DNA-Protein Interaction Deduced from *in vitro* DNA Evolution

Benoit Dubertret,[1] Shumo Liu,[2] Qi Ouyang,[2] and Albert Libchaber[1,2]
[1]*Center for Studies in Physics and Biology, The Rockefeller University, New York, New York 10021*
[2]*NEC Research Institute, 4 Independence Way, Princeton, New Jersey 08540*
(Received 19 January 2001)

We report the first experimental study of the dynamics of *in vitro* evolution of DNA. Starting from a random pool of DNA sequences, we used cycles of selection (binding to the *lac* repressor protein), amplification, and mutations to evolve to a unique DNA sequence, the *lac* operator. Statistical analysis of the DNA sequences obtained during the cycles of evolution shows that the DNA bases are selected at different rates. The rates of selection provide a quantitative measure of the interaction of the DNA bases with the protein during the complex formation. A model reproduces the evolution dynamics of the DNA population but cannot give the fine structure of the DNA-protein interaction.

DOI: 10.1103/PhysRevLett.86.6022 PACS numbers: 87.15.He, 87.15.Cc, 87.14.Gg, 87.15.Aa

The concept of *in vitro* evolution introduced in biology by Spiegelman *et al.* [1] has been used recently to evolve populations of DNA [2], RNA [3], or proteins [4]. In the case of DNA evolution, during one evolution cycle, a DNA population is amplified with the introduction of mutations through polymerase chain reaction (PCR). The DNA sequences are then selected according to their binding affinity to a protein and used as the input for the next cycle of evolution [5]. For the first time, we address experimentally the issue of the dynamics of DNA *in vitro* evolution and we show that it can be used to delineate DNA-protein interactions as they bind together.

In this Letter, we study the evolution of a random pool of DNA sequences to a *lac* operator sequence [6]. The selection criteria are defined through binding of DNA to the *lac* repressor protein (*lac*R) [7], a part of the *lac* operon which provided the first model for gene regulation [8]. Our work is motivated by three questions: (i) Can we measure the specificity of the *lac* repressor? (ii) Can we deduce from the dynamics of the evolution some information about the interaction between the protein and the DNA? (iii) Can we correlate the dynamics of the evolution with some known structural data of the protein-DNA complex [9]? In short, we found that the specificity of the *lac* repressor is very high: only one sequence is selected after five cycles of evolution. It is a 20-base-long sequence (*lac*O) close to the wild-type *lac* operator $lacO_1$. *lac*O is a palindrome of the left half of $lacO_1$, which lacks the central base pair [10]. This confirms that *lac*O is the DNA sequence that binds with the best affinity to the protein *lac*R [11]. A statistical analysis of the DNA sequences during the beginning of the evolution shows that some bases evolve faster than others. Our results are in good agreement with the current structural data on the *lac*O/*lac*R complex [9]. *lac*R binds *lac*O with two identical protein domains, or "headpieces." Each domain binds to ten DNA base pairs. When the two domains bind DNA together, they bend the DNA with a 40° angle [12]. When only one domain binds, the DNA is undistorted [13]. We can extract from our analysis the symmetric structure of the interaction of the protein *lac*R with the DNA *lac*O as well as the bending of *lac*O when bound to *lac*R.

We designed a 70-base-long DNA containing a 30-base-long random sequence flanked by two primer sites, each 20 base long. The amplification is done with 25 cycles of PCR using the Taq Polymerase that replicates DNA strand with the usual mutation rate around 10^{-4} per cycle per base [14]. After agarose gel purification, the amplified DNA is introduced in a tube coated with the *lac* repressor protein, *lac*R, prepared in-house. The LacR coated tube is obtained by incubation of a biotinylated lacR (using PinPoint *in vivo* biotinylation vector, Promega) in a streptavidin coated PCR tube (Boehringer Mannheim). Following 1 h incubation, the unbound DNA solution was washed out of the tube. After six washes with PBS, the DNA that was bound (specifically or not) to the protein was amplified by PCR, gel purified, and used in the next cycle of evolution. At each cycle, DNA molecules were subcloned in a plasmid (pBS, Stratagene), which was transfected in a bacteria and subsequently sequenced. Six to 17 sequences were obtained for each evolution cycle. The four relevant experimental parameters are the length of the random sequence, the starting concentration, the rate of mutation, and how extensive the wash is. We chose a 30-base-long random sequence, nine bases larger than the wild-type $lacO_1$, so that several *lac* operators could be selected. The starting number of DNA copies, 10^{-16} mole, is chosen such that there is a very low probability (less than 10^{-3}) that the starting pool contains a *lac* operator sequence. In this case, mutations are needed for the system to find its solution. The strength of the wash has been tuned so that there is around 10^{-17} mole of DNA (estimated with PCR) left in the tube after the wash.

To characterize the dynamics of the evolution, we measured at each cycle the distance and the phase between the 30-base-long evolving sequences and the 20-base-long fixed *lac*O sequence. The distance is defined as the smallest Hamming distance of all possible pairwise alignments

 0031-9007/01/86(26)/6022(4)$15.00

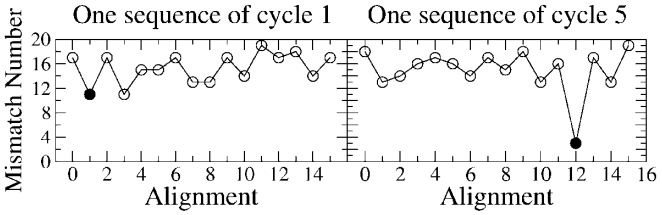

FIG. 1. Number of mismatches computed in all possible alignments between an evolving sequence and *lac*O. Alignment 0 means that the left of the two sequences coincide. Alignment 1 means that the 20-base-long sequence is shifted to the right by one base, and so on. The dark circles represent the first alignment with the smallest number of mismatches. It defines the distance and the phase of the evolving sequence.

between the two sequences. This distance is obtained for a shift between the two sequences that defines the phase of the evolving sequence (Fig. 1).

The dynamics of the evolution is characterized by both the sequence (Table I) and the phase (Table II). In the beginning (cycles 1 and 2), the average distance between *lac*O and the evolving sequences is around 12, a value typical of random sequences. At this point of the evolution, no *lac*O sequence appears and the phases are random. During cycles 3 and 4, the phases are still random, but the population of random sequences coexists with a population of sequences that is closer to *lac*O (distances of 7, 8, or less than 3). Good binding sequences have been found and are now enriched faster than random sequences. In cycle 3, the sequence with eight mismatches is very close to the right part of *lac*O, suggesting that the half operator is selected first. In cycle 4, the sequence seven mismatches away from *lac*O is three mismatches away from one of the wild-type $lacO_1$, which binds *in vitro* with 10 times less affinity than *lac*O [11]. This shows that at the initial stage, several sequences may be selected besides the best binding sequence. After cycle 5, although the sequences have converged to a unique point, three phases coexist: one on the left of the random sequence (phases 0 and 1), one in the right (phases 7 and 8), and one on the far right (phases 12 and 13) that partially overlap with the right primer. Only the left and the right phase remain after cycle 7. The third phase disappears: the sequences associated with this phase cannot converge completely to *lac*O since some bases are overlapping the primer site which does not mutate.

More information can be extracted from the evolution sequences. Before a phase has been chosen, the protein will bind to the DNA in all possible phases. Although the binding is mainly nonspecific, the DNA-protein interaction will be stronger for certain DNA bases than for others. These bases will be selected first and will thus appear with higher frequency when all possible alignments are analyzed. We compared each sequenced oligonucleotide with the *lac*O sequence in all possible phases. For each phase, we constructed the vector $(V_j^{\text{phase}})_{j=1,20}$, such that $V_j^{\text{phase}} = 1$ if the jth base of the two sequences is identical, $V_j^{\text{phase}} = 0$ otherwise. The sum over all phases, $\sum_{\text{phase}} V_j^{\text{phase}}$, gives the vector associated to a sequence. The vectors of all sequences belonging to the same cycle are then summed to yield the vector $(R_i^j)_{j=1,20}$, which characterizes the interaction of the *lac* repressor with the DNA during the ith cycle of the evolution. Each component R_i^j of R_i contains the number of times the jth base of the *lac*O was found at the right place in any possible phase in the sequences of the ith cycle. To take into account the inhomogeneities of the randomly generated sequences at cycle i (in commercially available pools of random sequences, there is typically more G and C than A or T), each component R_i^j is pondered by $1/\alpha_i^j$, where α_i^j is the total number of base j in cycle i ($j = A$, G, C, or T is the base associated with R_i^j). Each R_i is then normalized so that $\sum_{j=1,20} R_i^j = 1$. We performed the same statistical analysis comparing each sequenced oligonucleotide with the left-half and the right-half sequence (ten bases each) to obtain R_i^{left} and R_i^{right}.

The sum of the vectors R_i, R_i^{left}, and R_i^{right} for the four first cycles of evolution (a total of 45 sequences) is represented in Fig. 2. For the half operators, it appears that bases 6(T), 8(A), and 10(C) of the left half and bases 1(G) and 5(A) of the right half are essential for the interaction with one protein headpiece. These findings are in general agreement with the recent structural studies [15]. Conversely, bases 7(G) and 9(G) of the left half and bases 2(C) and 4(C) of the right half interact only weakly with *lac*R. Interestingly, if the wild-type 21-base-long *lac* operator, $lacO_1$, is viewed as two halves, reflected about a central base, left-side to right-side differences occur at bases 7(G) vs 15(A) and 9(G) vs 13(A), which correspond to the four

TABLE I. Number of sequences with m mismatches for each cycle of evolution.

Mismatches	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cycle 1												1	3	2		
Cycle 2											1	1	5		2	
Cycle 3									1			3	4	5		
Cycle 4			2	2				1				3	3	6		
Cycle 5			6	4	1											
Cycle 6	1		5	2												
Cycle 7		1	3	2												
Cycle 8		2	3	5												

TABLE II. Number of sequences with phase p. The phase corresponds to the alignment with the smallest distance reported Table I.

Phase	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Cycle 1		1			1		1		1	2						
Cycle 2	1	1		1		3	1		1							1
Cycle 3	1	1	1	1	3	2		1		1			1	1		
Cycle 4	4		2	1	2		1	1	2	1		1	2			
Cycle 5	3	1						3	1				1	2		
Cycle 6	2		1					3	1			1				
Cycle 7	2					1		3								
Cycle 8	3						1	4	2							

above mentioned bases interacting weakly with *lac*R. In the case of the whole *lac*O, among the three bases with the lowest frequencies, two (bases 3 and 12) do not appear to be important for the final *lac*O-*lac*R complex formation: after cycle 4, these bases are different than the *lac*O base in 25% of the DNA sequenced. On the contrary, base 10, which has the lowest frequency is selected in all the oligonucleotides sequenced after cycle 4. This observation suggests that bases that are secondary for the specific binding with the protein at the beginning of the interaction may become important for the interaction with the protein once the complex is formed. If the frequency

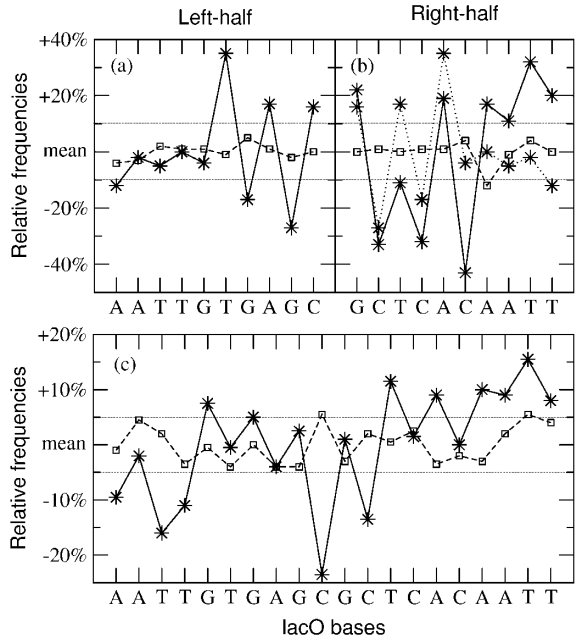

FIG. 2. Results of the sequence analysis for (a) the left-half *lac*O, (b) the right-half *lac*O and the inverted left-half *lac*O (dotted line), and (c) the whole *lac*O. ∗: evolving sequences for cycles 1–4; □: randomly generated sequences. In each case, to check the significance of our analysis, we computed the base frequency obtained comparing *lac*O with 45 random sequences. For the random sequences, all base frequencies are within 5% of the mean frequency (10% for the half *lac*O), in what we define as the noise region (represented by horizontal dotted lines). In the case of the evolution sequences, bases above that region are selected faster than average, denoting strong interaction with *lac*R at the beginning of the evolution process, bases that are selected slower than average (weak interaction) are under it.

distribution of the bases is to characterize their interaction with the protein, we expect that known structural properties of the *lac*R protein will be readily visible in the base frequency distributions of Fig. 2. In particular, the two half operators which interact with the protein through identical headpieces should show similar interactions. Indeed, we observe that the left half (close to the left primer) and the right half-*lac*O (close to the right primer) have similar frequency distributions, as shown in Fig. 2b. The discrepancies observed for the last four bases *AATT* are probably due to edge effects resulting from the interaction of the protein with the fixed primer. We also expect that the frequency distribution of the two half-*lac*O would be similar, respectively, to the left and to the right part of the frequency distribution of the whole *lac*O. Surprisingly, we found that only the distributions of the right halves show similarities. The left halves, on the contrary, have very different distributions, suggesting that the left part of the whole *lac*O interacts differently with the protein than the left half-*lac*O alone. In the case of *lac*O, the sum of the base frequencies of the right half is 10% higher than the one of the left half, suggesting that the right half operator has been selected before the left one. Once the right has been selected, the protein has to bend the DNA strongly to bind the second half. The difference observed for the left half is consistent with the conformational change (and thus the interaction changes) that occurs upon binding of *lac*R to the full *lac*O [12].

To better grasp the influence of the different parameters in our experiment, we developed a numerical simulation that models the *in vitro* evolution of DNA. In the simulation, the interaction between the DNA and the protein is approximated by a diagonal selection matrix. Thus, this simulation cannot give the fine structure of the DNA-protein interaction. Let us consider an initial random distribution, $\mathcal{N}^0$, of 30-base-long DNA sequences. In this starting pool, there are N_k^0 sequences that are k mismatches away from *lac*O in one of the ten possible phases. N_k^0 is proportional to $10 \times \binom{20}{k} \times 3^k/4^{20}$. This starting distribution $\mathcal{N}^0$ is then amplified with PCR. The amplification is responsible for the introduction of mutations at a rate r per cycle per base. During one cycle of PCR, a sequence S that is, say, i mismatches away from the operator undergoes m mutations. Among those m mutations, k occur on a base pair that was a mismatch, $m - k$ on one that was

matching the correct sequence (harmful mutation). Out of the k mutations, say l are good mutations and $k - l$ are neutral. $0 \le m \le 20$; $0 \le k \le m$; $0 \le l \le k$.

The probability $P(i, i + m - k - l)$ that S comes to $i + m - k - l$ mismatches from the *lac* operator is given by

$$\binom{30 - i}{m - k}\left(\frac{2}{3}\right)^{k-l}\binom{i}{k - l}\left(\frac{1}{3}\right)^{l}\binom{i - k + l}{l}r^{m}(1 - r)^{30-m}.$$

With $P(i, i + m - k - l)$ we define the PCR matrix

$$B_{ej} = \sum_{\substack{m,k,l \\ e+m-k-l=j}} P(e, e + m - k - l).$$

B_{ej} is the probability that a sequence e mismatches away from the operator comes to j mismatches in one cycle of PCR.

The vector $\mathcal{N}^1$, result of one cycle of PCR, is given by

$$N_e^1 = \sum_{k=1}^{20} B_{ek} N_k^0 .$$

The output of the 25 cycles of amplification is then multiplied by a diagonal selection matrix: $S_{ii} = 6^{-i}$, for $0 < i \le 5$ and $S_{ii} = 6^{-5}$ for $i > 5$. The choice of these numerical values is guided by the studies on the binding specificity of the *lac* operator/*lac* repressor complex [16]. The selection matrix rescales the total number of DNA sequences from an input of 10^{-10} mole to an output of 10^{-14} mole. Similarly, the PCR matrix amplifies the number of sequences back to 10^{-10} mole. Cycles of evolution are iterated until the number of *lac*O with one mismatch is 10 times larger than the number of perfect *lac*O.

Figure 3c gives the results of an evolution simulation with 10^{-16} mole of a starting random pool of DNA sequences, and a rate of mutation of 10^{-4} per cycle of PCR, per nucleotide. These values match our experimental parameters. The simulation shows that two populations of sequences are present during the evolution. As in our experiment, a unique population (population I) of DNA sequences centered around 15 mismatches is gradually depleted at the advantage of an other population (population II) of sequences with a very small number of mismatches. The convergence of the pool of random sequences appears to be a very robust characteristic of the system (Figs. 3a and 3b).

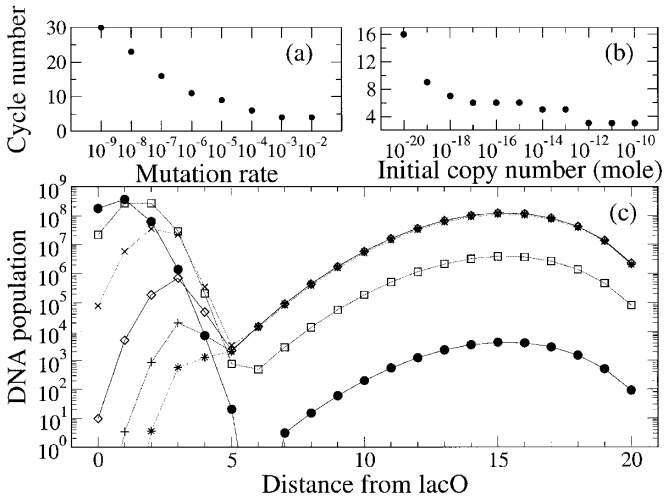

FIG. 3. (c) Numerical simulation of the DNA population evolution. Cycles 1 through 6 are represented, respectively, by $*$, $+$, $\diamond$, $\times$, $\square$, $\bullet$. Influence of (a) the rate of mutation per cycle per nucleotide and (b) the initial concentration on the speed of convergence. In (a) and (c), the initial number of DNA molecules is 10^{-16} mole; in (b) and (c) the rate of mutation is 10^{-4}.

We have shown that statistical analysis of the DNA sequences during DNA *in vitro* evolution is a method to analyze DNA-protein interaction. We were able to deduce the salient features of the LacR-LacO interactions. The small number of sequences analyzed limited our "resolution." The development of 2D single molecule PCR should provide a way to sequence rapidly several fold more sequences than here. The method we presented could then be a complement to structural biology and NMR analysis.

We thank M. E. Hatten, R. Bhatt, T. Tomoda, M. Goulian, D. Chatenay, and H. J. Bussemaker. This work was supported in part by the Mathers Foundation.

[1] D. R. Mills, R. L. Peterson, and S. Spiegelman, Proc. Natl. Acad. Sci. U.S.A. **58**, 217 (1967).
[2] Y.-Y. He, P. G. Stockley, and L. Gold, J. Mol. Biol. **255**, 55 (1996).
[3] C. Tuerk and L. Gold, Science **249**, 505 (1990).
[4] N. M. Low, P. Holliger, and G. Winter, J. Mol. Biol. **260**, 359 (1996).
[5] R. Pollock and R. Treisman, Nucleic Acids Res. **18**, 6197 (1990).
[6] W. Gilbert and A. Maxam, Proc. Natl. Acad. Sci. U.S.A. **70**, 3581 (1973).
[7] W. Gilbert and B. Müller-Hill, Proc. Natl. Acad. Sci. U.S.A. **56**, 1891 (1966).
[8] F. Jacob and J. Monod, J. Mol. Biol. **3**, 318 (1961).
[9] C. E. Bell and M. Lewis, Nat. Struct. Biol. **7**, 209 (2000).
[10] J. R. Sadler, H. Sasmor, and J. L. Betz, Proc. Natl. Acad. Sci. U.S.A. **80**, 6785 (1983).
[11] N. Lehming *et al.,* EMBO J. **6**, 3145 (1987).
[12] M. Lewis *et al.,* Science **271**, 1247 (1996).
[13] V. P. Chuprina *et al.,* J. Mol. Biol. **234**, 446 (1993).
[14] K. A. Eckert and T. Kunkel, Nucleic Acids Res. **18**, 3739 (1990).
[15] C. A. E. M. Spronk *et al.,* Structure **7**, 1483 (1999).
[16] J. L. Betz *et al.,* Gene **50**, 123 (1986).