Inference of DNA Sequences from Mechanical Unzipping: An Ideal-Case Study

V. Baldazzi,[1,2,3] S. Cocco,[2] E. Marinari,[4] and R. Monasson[3]

[1]*Dipartimento di Fisica, Università di Roma Tor Vergata, Roma, Italy*
[2]*CNRS-Laboratoire de Physique Statistique de l'ENS, 24 rue Lhomond, 75005 Paris, France*
[3]*CNRS-Laboratoire de Physique Théorique de l'ENS, 24 rue Lhomond, 75005 Paris, France*
[4]*Dipartimento di Fisica and INFN, Università di Roma La Sapienza, Piazza le Aldo Moro 2, 00185 Roma, Italy*

(Received 3 June 2005; published 30 March 2006)

The performances of Bayesian inference to predict the sequence of DNA molecules from fixed-force unzipping experiments are investigated. We show that the probability of misprediction decreases exponentially with the amount of collected data. The decay rate is calculated as a function of biochemical parameters (binding free energies), the sequence content, the applied force, the elastic properties of a DNA single strand, and time resolution.

DOI: 10.1103/PhysRevLett.96.128102 PACS numbers: 87.14.Gg, 05.10.−a, 87.15.Aa

DNA molecules are the support for genetic information, and knowledge of their sequences is essential from the biological and medical points of view. State-of-the-art DNA sequencing methods rely on biochemical and gel electrophoresis techniques [1] and are able to correctly predict about 99.9% of the bases. They were massively used over the past ten years to obtain the human genome (and those of other organisms). Nevertheless, the quest for alternative (cheaper and/or faster) sequencing methods is an active field of research. In this regard, single molecule micromanipulations are of particular interest. Among them are DNA unzipping under a mechanical action [2–6] or due to translocation through nanopores [7], the observation of the sequence-dependent activity of an exonuclease [8,9], the optical analysis of DNA polymerization in a nanochip device [10], and the detection of single DNA hybridization [11].

Hereafter, we focus on mechanical unzipping [Fig. 1(a)], first realized by Bockelmann and co-workers [2,3]. In their experiment, the strands are pulled apart under a constant velocity. The force is measured and fluctuates around 15 pN for the λ-phage DNA (a 48 502 base long virus), with higher (lower) values corresponding to the unzipping of GC (AT) rich regions. DNA (and RNA) molecules can also be unzipped under a constant force while measuring the distance between the open strands [4–7]. Various theoretical works have studied and reproduced the unzipping signal related to a given sequence [3,12–17]. We address here the inverse problem: Given an unzipping signal, can we predict the underlying DNA sequence? Based on a simple modeling of the unzipping dynamics which includes both thermal noise and the fluctuations of the open strands, we find that perfect prediction is possible provided enough experimental data are collected. The necessary amount of data is calculated as a function of the biochemical parameters (base pair energies), the force, the elastic properties of DNA single strands, and time resolution.

Let $b_n = \mathrm{A}$, T, C, or G denote the nth base along the $5' \rightarrow 3'$ strand (the other strand is complementary) and $B = \{b_1, b_2, \ldots, b_N\}$ the whole sequence. The free-energy excess when the first n base pair of the molecule are open with respect to the closed configuration is

$$G(n; B) = \sum_{m=1}^{n} g_0^{b_m, b_{m+1}} - n g_{ss}. \quad (1)$$

$g_0^{b_n, b_{n+1}}$ is the binding energy of base pair (bp) n [18]; it depends on the nearest bp due to stacking effects. g_{ss} is the work needed to stretch an open bp under a force f; according to the modified freely jointed chain model [12], $g_{ss} = 2f\ell \ln[\sinh(z)/z]/z$, where $z \equiv \ell_0 f/k_B T$, and $\ell_0 = 15$ Å and $\ell = 5.6$ Å are, respectively, the Kuhn and effective nucleotide lengths. The free-energy landscape $G(n; \Lambda)$ associated to the sequence $\Lambda = \{\lambda_1, \lambda_2, \ldots, \lambda_N\}$ of the

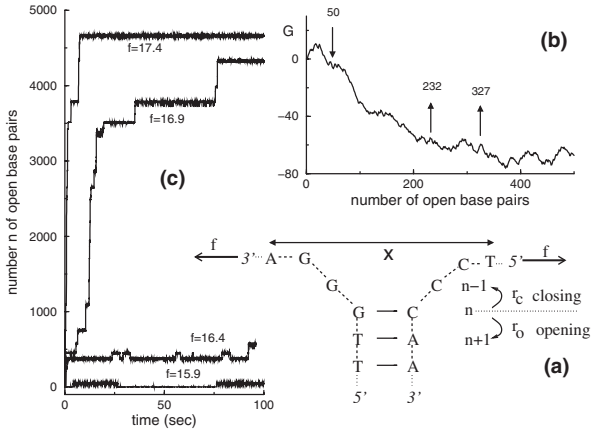

FIG. 1. Mechanical unzipping of a λ phage. (a) The extremities of the molecule are stretched apart under a force f while their distance x is measured. The fork at location n (number of open base pairs) moves backward or forward with rates r_c and r_o. (b) Free-energy landscape $G(n; \Lambda)$ versus n for the first 450 bases and force $f = 16.4$ pN. Down and up arrows indicate, respectively, a local minimum in $n = 50$ and two maxima in $n = 232$ and $n = 327$. (c) Time traces of n (for forces ranging from 15.9 to 17.4 pN) show a stick-slip behavior: Plateaus correspond to the deep minima of G and are separated by rapid unzipping jumps.

phage is shown in Fig. 1(b). We model the unzipping of the molecule as a one-dimensional biased random walk for the fork position (number of open bp) n in this landscape. Elementary opening ($n \to n+1$) and closing ($n \to n-1$) transitions happen with probabilities $r_o^{b_n,b_{n+1}} = r\exp\{g_0^{b_n,b_{n+1}}\}$, $r_c = r\exp\{g_{ss}\}$ per unit of time, respectively [Fig. 1(a)]. This choice for the rates has been shown [13] to reproduce quantitatively the behavior of unzipping experiments on short polynucleotides [4], with a typical frequency $r \simeq 10^{6-7}$ sec^{-1}. We use Monte Carlo (MC) dynamics to simulate this unzipping process. The output is a time trace $T = \{n_i, i = 1, 2, \ldots\}$ of the fork positions at discrete time $i \times \Delta$, where Δ is the inverse bandwidth; see Fig. 1(c).

We start by preprocessing our time trace T to obtain a set S whose $N \times N$ elements $S_{n,n'}$ are the numbers of transitions $n \to n'$, i.e., of pairs $(n_i, n_{i+1}) = (n, n')$ in T [19]. As our model is Markovian, the knowledge of higher order transitions, e.g., $n \to n' \to n''$, does not convey any further information. For this reason, S is called signal in the following. Information can be accumulated through repeated unzippings. Given R time traces T_j, $j = 1, 2, \ldots, R$, the total signal S is simply the sum of each one of the signals S_j.

Our dynamical model implicitly defines the distribution $\mathcal{P}(S|B)$ of signals given the sequence. The inverse problem, i.e., the prediction of the sequence given some signal, can be addressed within the Bayesian inference framework [20]. The probability that the DNA sequence is B given an observed S is

$$\mathcal{P}(B|S) = \frac{\mathcal{P}(S|B)\mathcal{P}_0(B)}{\mathcal{P}(S)}. \tag{2}$$

The value of B that maximizes this probability, $B^*(S)$, is our prediction for the sequence. In the absence of any *a priori* information about the sequence, $\mathcal{P}_0(B)$ is the flat distribution, equal to 4^{-N}. The maximization of $\mathcal{P}(B|S)$ over B then reduces to that of $\mathcal{P}(S|B)$.

As the signal is stochastic, the most likely sequence is not necessarily the correct one. Fortunately, errors are highly unlikely when R is large. More precisely, the rate of predicting some wrong sequence $B \neq \Lambda$

$$\epsilon_B = \text{probability}[B^*(S) = B] \sim e^{-R/R_B}, \tag{3}$$

where the probability is taken over the distribution $\mathcal{P}(S|\Lambda)$ of signals, decreases exponentially fast with R. R_B in (3) is interpreted as the (minimal) number of unzippings necessary to discard B from the candidates to the true sequence. Its value depends on the force, the bandwidth, the phage sequence, etc. Scaling (3) can be understood from a simple concentration argument. Let B be a sequence distinct from Λ. Figure 2 symbolizes the signal distributions associated to B and Λ. If the measured signal S is such that $\mathcal{P}(S|\Lambda) < \mathcal{P}(S|B)$, we erroneously infer that the true sequence is B. The probability of such a misprediction ϵ_B is given by the dark area in Fig. 2. The central limit theorem teaches us

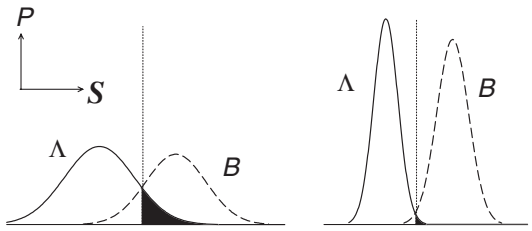

FIG. 2. Signal distributions $\mathcal{P}$ for the sequences Λ and B; the horizontal axis symbolizes the N^2-dimensional space of signals S. When the number R of unzippings increases (from left to right), distributions become more and more concentrated. Vertical dotted lines mark the sets of signals having equal probabilities with both distributions; the dark area is the probability ϵ_B (3) that Bayesian prediction is erroneous.

that, as the number R of unzippings increases (left to right in Fig. 2), the distributions become more and more peaked with standard deviation $\sigma \sim 1/\sqrt{R}$. The error ϵ_B is given by the Gaussian tail $\exp(-c_B/(2\sigma^2))$ and decreases exponentially fast with R; see the area shrinking in Fig. 2. Hence, (3) with $R_B = 2/c_B$. R_B can be calculated as a function of the sequence B with large deviations techniques. Because of space limitations, we do not give our expression for R_B in full generality but restrict to the two limiting cases of high and low bandwidths.

In the ideal case of infinite bandwidth ($\Delta \to 0$), jumps $|n_{i+1} - n_i|$ by more than one bp between two measures are very unlikely. Nonzero transition numbers are $S_{n,n} = t_n/\Delta$, where t_n is the time spent by the fork on base n and the numbers $S_{n,n\pm1}$ are openings and closings of the nth bp. The probability of this signal reads, up to a sequence-independent factor,

$$\mathcal{P}(S|B) \propto \prod_n (r_o^{b_n,b_{n+1}})^{S_{n,n+1}} r_c^{S_{n,n-1}} e^{-t_n(r_o^{b_n,b_{n+1}} + r_c)}. \tag{4}$$

B^* is easily found using the Viterbi algorithm [20,21]. The procedure is equivalent to a zero temperature 4×4 transfer matrix technique exploiting the nearest-neighbor nature of couplings between bases in (4). In practice, we first generate R unzipping signals by a MC procedure on the phage sequence. Then we use the Viterbi procedure (which ignores the phage sequence) to make a prediction for the sequence B^*. The quality of prediction on base n is estimated through the failure rate $\epsilon_n = \text{probability}[b_n^* \neq \lambda_n]$, where the probability is computed over different MC runs, each including R unzippings. Figure 3 shows the error ϵ_n for $R = 1$ (continuous curve) for the first 450 bases at a force of 16.4 pN. Values range from 0 (perfect prediction) to 0.75 (random guess of one among four bases). Comparison with Fig. 1(b) shows that ϵ_n is small in the flattest part of the free-energy landscape ($350 < n < 450$) or in local minima, e.g., the $n = 50$ base preceded by 4 weak bases and followed by 4 strong bases (...TTTA-A-GGCG...). Conversely, bases that are not well determined correspond to local maxima of the landscape, e.g.,

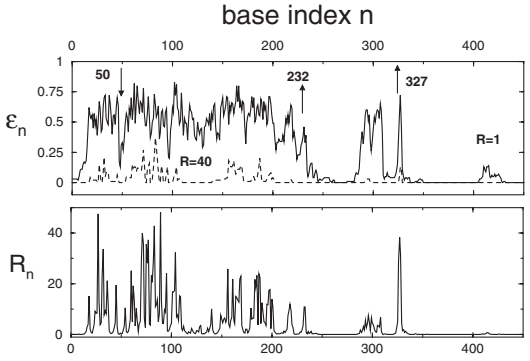

FIG. 3. Top: Probability ϵ_n of an error vs n, for the first 450 bp of the λ phage at $f = 16.4$ pN. The solid lines correspond to $R = 1$ unzipping, dotted lines to $R = 40$. Bottom: Theoretical values for the decay constants R_n. For instance, base 232 (arrow) is characterized by $R_{232} \simeq 10$ and is not (respectively, well) predicted with $R = 1$ (respectively, $R = 40$) unzippings.

bases $n = 327, 328$ between 7 strong and 7 weak bases (...GCCGCCG-TC-ATAAAAT...).

Performances are greatly improved by collecting information from multiple unzippings. Figure 3 (dashed curve) shows the drop in ϵ_n when R increases from 1 to 40; the intensity of the decay depends on the base index n. From (3), we indeed expect $\epsilon_n = e^{-R/R_n}$ at large R [22], where R_n is the maximum of R_B over the sequences B with $b_n \neq \lambda_n$. The exact expression for R_n is involved [21] but is accurately approximated by

$$R_n \simeq \frac{8}{\upsilon_n} \times \max_{b \neq \lambda_n} [(\Gamma^{\lambda_{n-1},b}_{\lambda_{n-1},\lambda_n})^2 + (\Gamma^{b,\lambda_{n+1}}_{\lambda_n,\lambda_{n+1}})^2]^{-1}, \qquad (5)$$

with $\Gamma^{x,z}_{x,y} = g_0^{x,z} - g_0^{x,y}$; υ_n is the average number of visits to bp n (transitions $n \longrightarrow n+1$) during one unzipping and is calculated from transient random walk theory [21]. Theoretical values for R_n from (5) are shown in Fig. 3. R_n varies from 0.1 to 45 with the base index n, in excellent agreement with the observed decay of ϵ_n between $R = 1$ and 40. The average fraction of mispredicted bases $\epsilon = \frac{1}{N}\sum_n \epsilon_n$ is a superposition of exponentials with n dependent decay constants and decreases very fast with R; see Fig. 4(a).

The dependence of ϵ upon the force and the biochemical parameters in Fig. 4 reflects the one of R_n. As the force f decreases, keeping the duration of the experiment fixed, fewer and fewer bases are unzipped, but these bases are visited more and more often by the fork [Fig. 1(c)]. These visits act as effective repetitions of the experiment; thus, the real number of unzippings necessary for good prediction, R_n, diminishes as $1/\upsilon_n$ [21]. Figure 4(a) confirms that the quality of prediction drastically improves (at fixed R) when f is lowered at a price of opening less and less bases in the same amount of time [Fig. 1(c)]. Most errors are due to the difficulty of distinguishing A from T and G from C. The probability that a weak (A or T) base is confused with

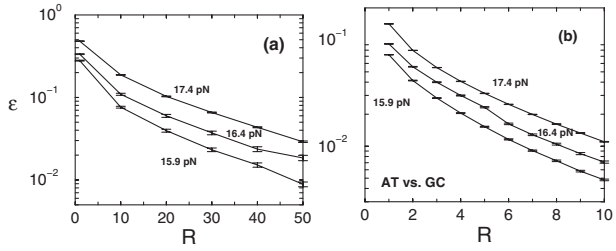

FIG. 4. (a) Fraction ϵ of mispredicted bases for the λ phage vs the number R of unzippings, averaged over 1000 samples of R unzippings, and for forces of 15.9, 16.4, and 17.4 pN (from bottom to top). (b) Same as (a), but we discriminate only among weak and strong basis.

a strong one (G or C), or vice versa, is plotted in Fig. 4(b). The decrease of the error ϵ with R is much faster for AT vs GC [Fig. 4(b)] than for complete [Fig. 4(a)] recognition. This can be explained from the scaling of R_n with the free-energy difference Γ between the bases to be recognized (5). We have $\Gamma \simeq 2.8$ to distinguish weak (AT or TA) from strong (CG or GC) bp [18], while complete recognition between the 4 bp types rather corresponds to $\Gamma \simeq 0.5$. The decay constants are, respectively, $R_n \simeq 1$ and $R_n \simeq 30$, in agreement with Figs. 4(b) and 4(a).

In practice, the bandwidth is limited by the viscous friction on the bead, the stiffness of the trap and linkers, with a current value of $\sim$1 kHz [3,23]. The delay between measures $\Delta \sim 1$ ms is larger than the opening time τ of an AT bp estimated to a fraction of microseconds [13]. The displacement of the fork $D = |n_{i+1} - n_i| \sim \Delta/\tau \sim 1000$ is well above the high bandwidth limit $D = 1$, previously studied. From the general concentration argument of Fig. 2, perfect sequence prediction is still possible but requires more unzippings. Calculations confirm that R_n is a growing function of Δ [21], with $R_n \propto \Delta$ in the low bandwidth limit $\Delta \gg \tau$.

This low bandwidth case, the hardest one as far as inference is concerned, can be understood in the case of short molecules, say, $\sim$10 bp. Then $\Delta = 1$ ms is of the order of the equilibration time of the random walk in the landscape of Fig. 1(b). The probability of the $n \longrightarrow n'$ transition in a time trace is essentially independent of n, and the probability of the signal S reads

$$\mathcal{P}(S|B) = \frac{(Rt/\Delta)!}{\prod_n U_n!} \prod_n p_B(n)^{U_n}, \qquad U_n \equiv \sum_m S_{m,n}. \quad (6)$$

Here $p_B(n)$ is the equilibrium measure with free energy (1) and t the duration of one unzipping. We find that ϵ_B obeys the scaling (3) with

$$R_B = \frac{\Delta}{t} \times \max_{0<\mu<1} \left| \ln \sum_n p_\Lambda(n)^\mu p_B(n)^{1-\mu} \right|^{-1}. \quad (7)$$

We have considered the sequence $\Lambda^{(30)}$ made of the first $N = 30$ bp of the phage at its equilibrium force, $f_c = 17$ pN. The most dangerous sequence B, i.e., the one

with the largest R_B, is obtained from $\Lambda^{(30)}$ through a mutation CG → GC of the 15th and 16th bp, with the result $R_B \times t \simeq 20$ s. In other words, $\Lambda^{(30)}$ can be recognized provided the time of the experiment is larger than 1 min.

Checking this prediction numerically is harder than in the high bandwidth case. Indeed, no exact and fast (running in a time growing polynomially with N and D) algorithm is known to infer the most likely sequence in the generic case of jumps of range D. We have extended the $D = 1$ Viterbi procedure through the use of a $4^D \times 4^D$ transfer matrix [21]. Our algorithm has an execution time linear in N but exponential in D and allows us to infer successfully sequences up to $D \sim 10$ only. For the $\Lambda^{(30)}$ example above, where $D = 30$, we have built a search algorithm, working in polynomial time but not guaranteed to find the most likely sequence. Nevertheless, $\Lambda^{(30)}$ is perfectly recovered with 100 s time traces, in very good agreement with theory.

Real experiments give access to the extension x of the open DNA (ssDNA) strands (with <1 nm resolution [24]) and not to the number n of open bp [Fig. 1(a)]. Each strand is made of n monomers modeled as springs with stiffness constant $K \simeq 170$ pN/nm at $f = 16$ pN and room temperature [12]. The distribution $A(x|n)$ of the extension x for a given n is roughly Gaussian, with mean nx_0, where $x_0 = dg_{ss}/df \simeq 0.9$ nm is twice the average extension of a ssDNA monomer, and standard deviation $\sqrt{n}\delta x$, where $\delta x = \sqrt{2k_BT/K} \simeq 0.2$ nm [25]. With Rouse dynamics [17], the longest relaxation time is, denoting the viscosity of the solvent by ζ, $t_r(n) \sim \zeta/(K\pi^2) \times (2n)^2 \sim 100n^2$ ps. ssDNA reaches equilibrium between two measures when $t_r(n) < \Delta$. Thus, for molecules with <100 bp, the stochastic evolution for x is Markovian for delays of the order of τ and above. The information in a time trace of the extension $\{x_i, i = 1, 2, \ldots\}$ is contained in the density $S(x, x')dxdx'$ of transitions from $x_i \in [x, x + dx] \to x_{i+1} \in [x', x' + dx']$. Our concentration argument of Fig. 2 can be repeated with $S_{n,n'}$ replaced with $S(x, x')$. Formula (3) for the probability of predicting sequence $B \neq \Lambda$ still holds but with a larger decay constant R_B^{ss} than in the absence of ssDNA fluctuations. In particular, the number of unzippings necessary for a reliable prediction of base n, $R_n^{ss} \simeq R_n \times (\delta x/x_0) \times \sqrt{n}$, increases as the square root of n whatever the delay Δ. The time required to correctly predict $\Lambda^{(30)}$ with a low bandwidth remains of the order of 1 min. In the high bandwidth limit, R_n^{ss} range from 10 to 100 for $n \leq 100$. Estimating the time for one unzipping to 1 s at 16.4 pN [Fig. 1(b)], it would take a few minutes to acquire the information; a shorter time would be found at lower temperatures. This time is not compatible with the large drift of simple optical trap setups (~10 nm/ min [3]) but requires the use of double optical traps, leading to a negligible drift <5 nm/h [26].

This work has been partially sponsored by the European EVERGROW (IST-001935) and STIPCO (HPRN-CT-2002-00319) programs and the French ACI-DRAB and PPF Biophysique-ENS actions.

[1] P. C. Turner *et al.*, *Molecular Biology* (Springer-Verlag, Berlin, 2000).
[2] B. Essevaz-Roulet, U. Bockelmann, and F. Heslot, Proc. Natl. Acad. Sci. U.S.A. **94**, 11 935 (1997).
[3] U. Bockelmann *et al.*, Biophys. J. **82**, 1537 (2002).
[4] J. Liphardt *et al.*, Science **292**, 733 (2001).
[5] S. Harlepp *et al.*, Eur. Phys. J. E **12**, 605 (2003).
[6] C. Danilowitcz *et al.*, Proc. Natl. Acad. Sci. U.S.A. **100**, 1694 (2003).
[7] A. F. Sauer-Budge *et al.*, Phys. Rev. Lett. **90**, 238101 (2003); J. Mathé *et al.*, Biophys. J. **87**, 3205 (2004).
[8] A. M. van Oijen *et al.*, Science **301**, 1235 (2003).
[9] T. Perkins *et al.*, Science **301**, 1914 (2003).
[10] M. J. Levene *et al.*, Science **299**, 682 (2003).
[11] M. Singh-Zocchi *et al.*, Proc. Natl. Acad. Sci. U.S.A. **100**, 7605 (2003).
[12] S. Cocco, R. Monasson, and J. Marko, C.R. Physique **3**, 569 (2002).
[13] S. Cocco, R. Monasson, and J. Marko, Eur. Phys. J. E **10**, 153 (2003).
[14] D. K. Lubensky and D. R. Nelson, Phys. Rev. Lett. **85**, 1572 (2000); Phys. Rev. E **65**, 031917 (2002).
[15] U. Gerland, R. Bundschuh, and T. Hwa, Biophys. J. **81**, 1324 (2001).
[16] M. Manosas and F. Ritort, Biophys. J. **88**, 3224 (2005).
[17] D. Marenduzzo *et al.*, Phys. Rev. Lett. **88**, 028102 (2002).
[18] M. Zuker, Curr. Opin. Struct. Biol. **10**, 303 (2000); from Santa Lucia, Jr. [Proc. Natl. Acad. Sci. U.S.A. **95**, 1460 (1998)], $g_0^{AA} = 1.78$, $g_0^{AT} = 1.55$, $g_0^{AC} = 2.52$, $g_0^{AG} = 2.22$, $g_0^{TA} = 1.06$, $g_0^{TC} = 2.28$, $g_0^{TG} = 2.54$, $g_0^{CC} = 3.14$, $g_0^{CG} = 3.85$, $g_0^{GC} = 3.90 k_BT$ at $T = 25$ C, 150 mM Na.
[19] The dynamical model authorizes only $n \to n \pm 1$ transitions in infinitesimal time, but generic transitions $n \to n'$ are allowed in a finite time Δ.
[20] D. J. C. McKay, *Information Theory, Inference and Learning Algorithms* (Cambridge University Press, Cambridge, England, 2003).
[21] V. Baldazzi *et al.* (to be published).
[22] It is assumed here that the first-passage time of the fork on base n is smaller than the duration t of each unzipping.
[23] B. Onoa *et al.*, Science **299**, 1892 (2003).
[24] K. C. Neumann and S. Block, Rev. Sci. Instrum. **75**, 2787 (2004).
[25] R. E. Thompson and E. D. Siggia, Europhys. Lett. **31**, 335 (1995).
[26] J. W. Shaevitz *et al.*, Nature (London) **426**, 684 (2003).