We introduce and test a method to predict the sequence of DNA molecules from *in silico* unzipping experiments. The method is based on Bayesian inference and on the Viterbi decoding algorithm. The probability of misprediction decreases exponentially with the number of unzippings, with a decay rate depending on the applied force and the sequence content.

DNA molecules are the support for the genetic information, and knowledge of their sequences is very important 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 year to obtain the human genome (and the ones of other organisms).

Nevertheless, the quest for alternative (cheaper and/or faster) sequencing methods is an active field of research. In this regard, recent single molecule manipulations are of particular interest. Among them are DNA unzipping under a mechanical action 2, 3, 4, or due to translocation through nanopores 5, the observation of the sequence-dependent activity of an exonuclease 6, 7, the optical analysis of DNA polymerization in a nano-chip device 8, and the detection of single DNA hybridization 9. Hereafter, we focus on mechanical unzipping (see Figure 1), first realized by Bockelmann, Heslot and coworkers in 1997 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 (respectively, lower) values corresponding to the unzipping of GC (AT) rich regions. Researchers have also unzipped RNA molecules 10, 11, or DNA under a constant force (instead of velocity) 12. Figure 2A sketches a fixed-force output signal, with its pauses in the opening at sequence-specific positions.

Various theoretical works have studied and reproduced the unzipping signal related to a given sequence 13, 14, 15, 16, 17. Hereafter we address the inverse problem: given an unzipping signal (for example the one of Figure 2B), can we predict the underlying sequence? We propose a Bayesian inference method to solve this problem 18, and test it *in silico* on the λ-phage. We analytically study the dependence of the quality of the prediction on the sequence content, on the force, and on the number of unzippings. Finally we list the main obstacles to be circumvented prior to practical applications.

Let $S = \{b_1, b_2, \ldots, b_N\}$ denote the sequence of $N$ bases along the $5' \rightarrow 3'$ strand (the other strand is complementary). We model the unzipping of the molecule through the evolution of the number $n$ of open base pairs 18: base pair opening ($n \rightarrow n + 1$) and closing ($n \rightarrow n - 1$) happen with rates (Figure 1)

$$r_o(n) = r \exp\{g_0(n)\}, \quad r_c = r \exp\{g_{ss}\}. \quad (1)$$

$g_0(n)$ is the binding energy of base pair (bp) $n$ in units of $k_B T$ 19; it depends on the base $b_n = A, T, G,$ or $C$ and, due to stacking effects, on the nearest base $b_{n+1}$. $g_{ss}$ is
the work needed to stretch an open bp under a force \( f \) in units of \( k_B T \); according to the modified freely–jointed–
chain model \( g_{ss} = -2T/\ell_0 \ln \sinh(x)/x \) where \( x \equiv \ell_0 f/k_BT \) and \( \ell_0 = 15 \) Å and \( \ell = 5.6 \) Å are, respectively, the Kuhn and effective nucleotide lengths. Relation (1) implies that the opening rate at base \( n \) is a function of the sequence, \( r_o(n) = r_o(b_n, b_{n+1}) \), while the closing rate \( r_c \) only depends on the force \( \lambda \). This a priori choice has been shown \( \textit{21} \) to reproduce quantitatively the behavior of unzipping experiments on short polynucleotides \( \textit{3} \), with a typical frequency \( \gamma \approx 10^{6−7} \) sec\(^{-1}\).

Rates \( \textit{1} \) define a one-dimensional biased random walk for the fork position (number of open bp) \( n(t) \) in the potential \( g(n) = n g_{ss} - \sum_{i=1}^{n} g_{i} (i) \), that can be interpreted as the free energy of the molecule when the first \( n \) bp are open. We show in Figure \( \textit{2 B} \)\&\( \textit{C} \) a typical time-trace of \( n(t) \) generated by Monte Carlo (MC) simulation for the \( \lambda \)-phage sequence, together with the free energy landscape \( g(n) \). Plateaus of \( n(t) \) coincide with deep local minima of \( g(n) \), where the fork remains trapped for a long time. As the force increases, opening becomes more favorable, and plateaus shrink.

Our \textit{in silico} time-traces are stochastic due to the thermal noise: two runs will give different traces. The probability of a time-trace only depends on the set \( \mathcal{N} = \{t_n, u_n, d_n\} \) of times \( t_n \) spent on each base \( n \), and of numbers \( u_n \) and \( d_n \) of up (\( n \rightarrow n + 1 \)) and down (\( n \rightarrow n − 1 \)) transitions respectively. Given the sequence \( S \), this probability reads

\[
P(\mathcal{N}|S) = c \prod_{n} M(b_n, b_{n+1}; t_n, u_n, d_n) ,
\]

where \( c \) is a (sequence-independent) normalization constant and \( M(b_n, b_{n+1}; t_n, u_n, d_n) = r_o(b_n, b_{n+1})^{u_n} r_c^{d_n} \exp\{-r_o(b_n, b_{n+1}) + r_c\} \). Equation (2) provides the solution of the direct problem: given the sequence \( S \) what is the distribution of the time-traces \( \mathcal{N} \)? The inverse problem, that is the prediction of the sequence given some time-trace, can be addressed within the Bayesian inference framework. The probability that DNA sequence is \( S \) given an observed \( \mathcal{N} \) is \( \textit{12} \)

\[
P(S|\mathcal{N}) = \frac{P(\mathcal{N}|S) P_0(S)}{P(\mathcal{N})} .
\]

The value of \( S \) that maximizes this probability, \( S^* \), is our prediction for the sequence. In the absence of any a priori information about the sequence, \( P_0(S) \) is the flat distribution, equal to \( 4^{-N} \). The maximization of \( P(S|\mathcal{N}) \) then reduces to that of \( P(\mathcal{N}|S) \) (Equation (2)).

In practice the most likely sequence \( S^* \) may be found using the Viterbi algorithm \( \textit{21} \). The procedure is equivalent to a zero temperature transfer matrix technique exploiting the nearest-neighbor nature of couplings between bases in \( \textit{1} \). The probability \( P_n \) for the base \( b_n \) fulfills the recursive equation

\[
P_{n+1}(b_{n+1}) \propto \max_{b_n} P_n(b_n) M(b_n, b_{n+1}; t_n, u_n, d_n) ,
\]

where the proportionality constant is irrelevant for our purpose. The maximum in \( \textit{3} \) is reached for some base \( b_{n+1}^{\text{max}}(b_n) \) that depends on the next base \( b_{n+1} \). Starting from \( P_1(b_1) = \frac{1}{2} \), we obtain the probability \( P_N(b_N) \) for the last base of the sequence through iterations of \( \textit{4} \). Maximization of \( P_N(b_N) \) yields the most likely value for this last base, \( b_N^* \). The whole optimal sequence \( S^* \) is then recursively obtained from the relation \( b_{n+1}^* = b_{n+1}^{\text{max}}(b_n^*) \).

We have tested our sequencing method on the \( \lambda \)-phage. First we build a dynamical process on the sequence \( S^\lambda \) of the phage with rates \( \textit{1} \), and generate an unzipping trace \( \mathcal{N} \) by a MC procedure. Then we use the Viterbi procedure (which ignores the phage sequence) to make a prediction for the sequence, \( S^* \), from this signal \( \mathcal{N} \). We estimate the error over the prediction about base \( n \) from the failure rate

\[
\epsilon_n = \text{Probability} \{ b_n^* \neq b_n \} ,
\]

where the probability is computed by repeating the procedure over different MC runs. The errors \( \epsilon_n \) are shown in Figure \( \textit{3} \) (with the continuous curve) for the first 450 bases at a force of 16 pN. Values range from 0 (perfect prediction) to 0.75 (random guess of one among four bases). A comparison with the free energy \( g(n) \) (Figure \( \textit{2} \) shows that \( \epsilon_n \) is small in the flattest part of the 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-GGCC...). Conversely, bases that are not well determined correspond to local maxima of the landscape e.g. \( n = 327, 328 \) bases between 7 strong and 7 weak bases (...GCCGCC-GTC-ATAAAAA...). We plot the average fraction of mispredicted bases, \( \epsilon = \frac{1}{N} \sum_2^{N} \epsilon_n \) in Figure \( \textit{4 A} \). As shown in Figure \( \textit{2} \), for a larger force, there are more open bases (about 60, 600 and 5000 at 15.5, 16 and 17 pN in about 100 seconds), but the time spent on each base is smaller, and therefore \( \epsilon \) is larger (\( \epsilon = 20\%, 23\%, 47\% \)). 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 a strong one (G or C), or vice-versa, is plotted in Figure \( \textit{4 B} \).

Performances can be greatly improved by collecting information from multiple unzippings. As the number of passages over the same base \( n \) gets larger, the total waiting times \( t_n \) and transition parameters \( u_n, d_n \) become less affected by fluctuations, and reflect more faithfully the thermodynamic signature of the base. In practice, we look for the most likely sequence \( S^* \) given \( R \) unzipping signals \( \mathcal{N}_1, \mathcal{N}_2, \ldots, \mathcal{N}_R \). Figures \( \textit{4 A} \) and \( \textit{4 B} \) shows the drop in the probability of error when the number \( R \) of unzippings increases. Observe from Figure \( \textit{3 A} \) that the decay of \( \epsilon_n \) with \( R \) \( \textit{4} \) varies from base to base. The
decrease of the total error \( \epsilon \) is much faster for AT vs. GC (Figure 4B) than for complete (Figure 4A) recognition.

It is useful to build indicators of performances that do not rely on the exact knowledge of the unzipped sequence (used here for checking the quality of our results but unknown in practical applications). To this aim, we calculate the optimal sequences \( S^*_n \) when base \( n \) is constrained to value \( b \), and the corresponding probabilities \( P^*_n(b) \). We then define the Shannon entropy

\[
\sigma_n = - \sum_{b=A,T,G,C} \langle P^*_n(b) \rangle \log_2 P^*_n(b) ,
\]

where \( \langle \cdot \rangle \) denotes the average over MC data. \( \sigma_n \) is low when one of the four bases has much higher probability than the other ones and close to unity for uncertain predictions (equiprobable bases). Figure 4 shows that \( \sigma_n \) and \( \epsilon_n \) as a function of the base index \( n \) are indeed very similar: the Shannon entropy is a good indicator of the success of our reconstruction.

Our analytical study of the dependence of the quality of the prediction upon the force, the sequence content, and the number of unzipping confirms that the probability of error \( \epsilon_n \) decreases very quickly with \( R \),

\[
\epsilon_n \sim e^{-R/R^c_n} .
\]

As \( f \) decreases to its critical value (below which the molecule cannot open), the decay constant \( R^c_n \) decreases to zero, and predictions drastically improve at fixed \( R \). Our theoretical values for \( R^c_n \) are shown in Figure 3B for \( f = 16 \) pN, and vary from 0.1 to 45 with the base index \( n \). The agreement with the decay of \( \epsilon_n \) from \( R = 1 \) to 40 unzippings (Figure 4A) is excellent. Note that \( \epsilon \) in Figure 4 is not a pure exponential, but a superposition of exponentials with \( n \)-dependent decay constants \( R^c_n \). We now present the calculation of \( R^c_n \) in three steps.

(a) Pairing only, high force. Assume first that there are only 2 and not 4 bp-types, called + and −, and no stacking interaction. Call \( \Delta \) the difference between the (pairing) free-energies of + and − bp, and \( \langle t_{\pm} \rangle \) the average time spent by the fork on a ± bp before moving forward or backward. Consider now a bp of type \( b \) and call \( t \) the time spent on this bp divided by the number \( R \) of unzippings. From the central limit theorem, for large \( R \), \( t \) gets narrowly peaked around its mean value \( \langle t_b \rangle \), with Gaussian fluctuations \( \delta t \sim R^{-\frac{1}{2}} \). Bayes prediction \( \epsilon \) will be erroneous, \( b^* = -b \), when \( t \) is closer to \( \langle t_{-b} \rangle \) than to its expected value \( \langle t_b \rangle \). The probability of error is thus given by the Gaussian tail, and scales as \( \epsilon \sim \exp(-\delta t^{-2}) \), hence \( \epsilon \). A careful calculation \( 20 \) gives the precise value of the decay constant in \( \epsilon \),

\[
R^c = \frac{1}{\tau - 1 - \ln \tau} \quad \text{with} \quad \tau = \frac{\Delta}{1 - e^{-\Delta}} .
\]

Good predictions are obtained when the molecule is unzipped a few \( R^c \) times (for example \( R \approx 4R^c \) gives \( \epsilon \approx 2\% \)). To distinguish weak (AT) from strong (CG) bp only we have \( \Delta \approx 2.8 \) \( 19 \), and \( R^c \) ≃ 1 (Figure 4B), while complete recognition corresponds to \( \Delta \approx 0.5 \) and \( R^c \approx 30 \) (Figure 4A).

(b) Pairing and Stacking, high force. In presence of stacking interactions, the error \( \epsilon_b \) on base \( b \) depends on the neighboring bases, \( x \) and \( y \). At large \( R \), errors are rare and are typically due to a single base mis-prediction e.g. \( b \to b' \). The probability \( \epsilon_{b\to b'} \) of this mistake is the product of the probabilities \( \epsilon_{xb\to x'b} \) and \( \epsilon_{by\to y'b} \) of the two bond violations. We estimate \( \epsilon_{xb\to x'b} \sim e^{-R/R^c_{xb\to x'b}} \) from \( 7 \) where \( R^c_{xb\to x'b} \) is given by \( 8 \) with \( \Delta = \delta_{b'b}^y - \delta_{b}^x \). A similar expression is readily obtained for the by bond. Knowing the asymptotic behavior of \( \epsilon_{b\to b'} \), we calculate \( \epsilon_b \sim e^{-R/R^c_{xb\to x'b}} \) by selecting the worst value for \( b' \),

\[
\frac{1}{R^c_{xb\to x'b}} = \min_{b'} \left[ \frac{1}{R^c_{xb\to x'b}} + \frac{1}{R^c_{by\to y'b}} \right] .
\]

The above derivation is confirmed by exact calculations based on techniques for 1D disordered systems \( 20, 22 \).
(c) Moderate force. The above calculations are correct for high forces. At moderate forces, bp can close and are visited several times by the fork. The effective number of unzippings is \( R \times \langle u_n \rangle \), where \( \langle u_n \rangle \) is the average number of openings of bp \( n \) during a single unzipping. The decay constant is thus, from 7, 

\[
R^c_n = R^c_{bn-1}b_n b_{n+1}/\langle u_n \rangle.
\]  

(10)

As the force is lowered, \( \langle u_n \rangle \) increases (from 1 at high force), and \( R^c_n \) diminishes. To calculate \( \langle u_n \rangle \), we consider the 1D transient random walk defined by the probabilities \( q_m = r_c/(r_a(m) + r_c) \) and 1 - \( q_m \) for closing or opening bp \( m \). Let \( p^{(n)}_m \) be the probability that the fork will never reach position \( m(> n) \). The ratio \( p^{(n)}_m = p^{(n)}_m/p^{(n)}_{m+1} \) fulfill the Riccati recursion relation 20. \( p^{(n)}_{m+1} = (1 - q_{m+1})/(1 - q_m p^{(n)}_m) \). Iterating with boundary condition \( p^{(n)}_m = 0 \) allows us to obtain \( \langle u_n \rangle = 1/p^{(n+1)}_n = \prod_{m>n} p^{(n)}_m \).

Finally we discuss the difficulties hindering a direct application of our inference method to real data (see also 21), and possible way-outs.

First, temporal resolution is limited in practice. The frequency bandwidth is controlled by the viscous friction and the stiffness of the setup, with a typical value of 10 kHz 3, 21. The corresponding time, \( \delta \tau \approx 100 \) \( \mu \)sec, is about 10 (resp. 200) times longer than the typical opening time for GC (resp. AT) bp. As a result, the fork can move by \( D(> 1) \) bp during the time interval \( \delta \tau \). We have taken into account such moves by considering interactions between bases at distance \( \leq D \) in the probability \( P(\mathcal{V}|S) \), and modified the reconstruction procedure accordingly (the transfer matrix has now dimension \( 4^D \) 20). In practice, when \( \delta \tau = 1 \) \( \mu \)sec, sequences cannot be predicted with the usual \( D = 1 \) reconstruction procedure, but are correctly inferred with the \( D = 6 \) procedure. Though time resolution is currently far below this limit, future experimental progresses, and new technologies e.g. combination of optical trap and single-molecule fluorescence 24, could help bridging the gap.

Secondly, thermal fluctuations of the open strands lead to an uncertainty \( \delta n \) over the position \( n \) of the fork 22. e.g. \( \delta n \approx 5 \) for \( f \approx 15 \) pN and \( n = 300 \) open bp. The presence of correlations between bases at distance \( D \leq \delta n \) does not affect the result 7 for \( c_m \) as long as the relaxation time of the strands is smaller than the bp opening time i.e. up to a few hundreds open bp. What happens for larger values of \( n \) is currently under study.

Thirdly, we have assumed so far to have a perfect knowledge of the dynamics of unzipping. In practice, any functional form for \( P(\mathcal{V}|S) \) will be only approximate for a given experimental setup. A possible way-out based on a learning principle is the following: in a first stage unzipping data corresponding to a known sequence (\( \lambda \)-phage) are collected to calibrate \( P \), in a second stage predictions are made for new sequences.

Last of all, our study of fixed-force unzipping shows that bases located in local minima of the free-energy landscape are well predicted, while maxima are much harder to predict. Accuracy could be greatly improved through an adequate force vs. time scheme capable of bringing the fork in the right place and making it spend time there. Investigation of the fixed-velocity case, where the force signal is remarkably affected by single base mutation 3, will be very interesting.

In conclusion, we hope the present study will motivate further work to assess and improve the performances of unzipping-based sequencing.

This work has been partially sponsored by the EC FP6 program under contract IST-001935, EVERGROW, and the French ACI-DRAB & PPF Biophysique-ENS actions.

\[\text{[1] P.C. Turner, A.G. McLennan, A.D. Bates, M.R.H. White, Molecular Biology, Springer-Verlag (2000).} \]
\[\text{[2] B. Essevaz-Roulet, U. Bockelmann, F. Heslot, \textit{Proc. Natl. Acad. Sci. (USA)} 94, 11935 (1997).} \]
\[\text{[3] U. Bockelmann et al. \textit{Biophys. J.} 82, 1537 (2002).} \]
\[\text{[4] J. Liphardt et al. \textit{Science} 297, 733 (2001).} \]
\[\text{[5] S. Harlepp et al. \textit{Eur. Phys. J. E} 12, 605 (2003).} \]
\[\text{[6] C. Damiłowicz et al. \textit{Proc. Natl. Acad. Sci. (USA)} 100, 1694 (2003).} \]
\[\text{[7] J. Mathé et al. \textit{Biophys. J.} 87, 3205 (2004).} \]
\[\text{[8] M. van Oijen et al. \textit{Science} 301, 124 (2003).} \]
\[\text{[9] T. Perkins et al. \textit{Science} 301, 1914 (2003).} \]
\[\text{[10] M.J. Levene et al. \textit{Science} 299, 682 (2003).} \]
\[\text{[11] M. Singh-Zocchi et al. \textit{Proc. Natl. Acad. Sci. (USA)} 100, 7605 (2003).} \]
\[\text{[12] S. Cocco, R. Monasson, J. Marko. \textit{C.R. Physique} 3, 569 (2002).} \]
\[\text{[13] S. Cocco, R. Monasson, J. Marko. \textit{Eur. Phys. J. E} 10, 153 (2003).} \]
\[\text{[14] D.K. Lubensky, D.R. Nelson. \textit{Phys. Rev. Lett.} 85, 1572 (2000); \textit{Phys. Rev. E} 65, 031917 (2002).} \]
\[\text{[15] U. Gerland, R. Bündschuh, T. Hwa. \textit{Biophys. J.} 81, 1324 (2001).} \]
\[\text{[16] M. Manosas, F. Ritort. \textit{cond-mat/0405037} (2004).} \]
\[\text{[17] D. Marenduzzo et al. \textit{Phys. Rev. Lett.} 88, 028102 (2002).} \]
\[\text{[18] D.H. DeGroot, Probability and Statistics, Addison-Wesley Publishing Co. (1986).} \]
\[\text{[19] M. Zuker, \textit{Curr. Opin. Struct. Biol.} 10, 303 (2000). From Santa Lucia Jr. Proc. Natl. Acad. Sci. (USA) 95, 1460 (1998). } \]
\[\text{[20] V. Baldazzi et al., in preparation (2005).} \]
\[\text{[21] A.J. Viterbi, \textit{IEEE Trans. Inf. Th.} 13, 260 (1967).} \]
\[\text{[22] F.J. Dyson \textit{Phys. Rev.} 92, 1331 (1953).} \]
\[\text{[23] R.E. Thompson, E.D. Siggia. \textit{Europhys. Lett.} 31, 335 (1995).} \]
\[\text{[24] B. Onoa et al. \textit{Science} 299, 1892 (2003) (supplementary materials).} \]
[25] U. Gerland, R. Bundschuh, T. Hwa. *Phys. Biol.* **1**, 19 (2004).

[26] M.J. Lang, P.M. Fordyce, S.M. Block. *J. Biol.* **2**, 6 (2003).