Statistical Inference for Nanopore Sequencing with a Biased Random Walk Model

Kevin J. Emmett,[†*] Jacob K. Rosenstein,[¶] Jan-Willem van de Meent,[‡] Ken L. Shepard,[§] and Chris H. Wiggins[‡]

[†]Department of Physics and [‡]Department of Applied Physics and Applied Math and [§]Department of Electrical Engineering, Columbia University, New York, New York; and [¶]School of Engineering, Brown University, Providence, Rhode Island

ABSTRACT Nanopore sequencing promises long read-lengths and single-molecule resolution, but the stochastic motion of the DNA molecule inside the pore is a current barrier to high accuracy reads. We develop a method of statistical inference that explicitly accounts for this error and demonstrate that high accuracy (>99%) sequence inference is feasible even under highly diffusive motion by using a hidden Markov model to jointly analyze multiple stochastic reads. Using this model, we place bounds on achievable inference accuracy under a range of experimental parameters.

Received for publication 31 July 2013 and in final form 31 July 2013.

*Correspondence: kje@phys.columbia.edu.

Rapid advances in DNA sequencing technologies have led to an explosion in available nucleotide sequence data, greatly enhancing our understanding of the genomic basis of many biological processes. However, the short length of the raw reads means high coverage is required for reliable sequence assembly. Nanopore sequencing has emerged as a candidate to supercede current generation sequencing and allow for theoretically unlimited read length (1). A number of strategies have been proposed, with the common basis of detecting individual nucleotides as they pass through a nanometer-scale aperture in a thin membrane separating two electrolytes (2). To date, a significant obstacle of nanopore approaches has been overcoming the fast stochastic motion of the individual molecules as they are driven through the pore (3, 5). Ideally, passage of DNA through the pore would be unidirectional and each base would have a well-resolved signal. Recent methods have demonstrated an ability to controllably 'ratchet' DNA molecules through a nanopore one base at a time, although motion of the molecule can still occur in both forward and backward directions within a single read (7, 8). Unidirectional motion remains difficult to reliably achieve, leading to a source of error in the read sequence recognized, but not previously addressed, by existing models.

In this letter, we analyze the effect of diffusive motion on achievable read accuracy and propose a statistical method to account this error. The method uses hidden Markov models (HMMs), which have recently been used to study multi-base resolution in a nanopore sequencer (10), but have not yet been applied to the problem of diffusive motion inside the pore. We show that combining multiple reads from an input DNA sequence allows accurate estimates of the sequence, both in the presence of highly diffusive molecular motion and high base-call error rates.

Assuming no DNA-pore interaction, polymer translocation can be modeled as one-dimensional diffusion with drift, where the probability of a displacement x in a time interval t is given by

$$p(x,t) = \frac{1}{\sqrt{4\pi D t}} \exp\left[-\frac{(x-vt)^2}{4Dt}\right]. \qquad (1)$$

The drift velocity $v = F/\gamma$ is determined by the driving force F and drag coefficient γ, which also determine the diffusion constant $D = \gamma\, k_B T$ via the fluctuation-dissipation theorem. A nondimensional forward bias is defined as $\xi = Fa/4kT$. We assume F is tuned to obtain an expected displacement $v\tau = a$, where a is one nucleotide distance and τ is the sampling interval. Defining a discretized sequence position $z = \mathrm{nint}(x/a)$, the probability of moving to position z_s given a previous position z_{s-1} is given by

$$p(z_s \mid z_{s-1}) = \int_{a(z_s - z_{s-1}) - a/2}^{a(z_s - z_{s-1}) + a/2} p(x,\tau)\, dx \qquad (2)$$

Given an input DNA sequence of length L, we generate a single output read by stepping through the sequence with transition probabilities $p(z_s \mid z_{s-1})$, at each step making a base-call with error probability ϵ, which is independent of error introduced due to backward motion. Finally, we assume an appropriate method of making a base-call from the raw signal, which has been computationally studied in (11). From this simple physical model, we generate a set of N simulated reads X_n, each having a unique length T_n. N is a notion of sequence coverage in terms of identical molecules sequenced, which could result from either amplification or resequencing of the initial molecule. A schematic representation of a nanopore sequencing device is shown in Fig. 1 *A*. The relationship between forward bias, drag coefficient, and sampling frequency required for single base resolution is shown in Fig. 1 *B*. In Fig. 1 *C* we plot $p(z_s \mid z_{s-1})$ at several different forward biases. In Fig. 1 *D* the distribution of read lengths at these forward biases is shown.

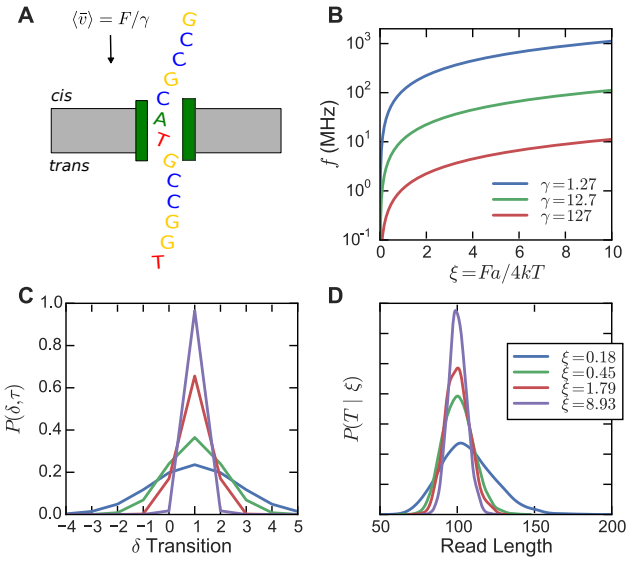

FIGURE 1 Random walk model of nanopore sequencing. (A) Schematic representation of an idealized nanopore sequencing device. (B) Minimum sampling frequency required for single base resolution as a function of forward bias, ξ, and drag coefficient, γ. (C) Transition probabilities and (D) read length distributions at different forward biases (sequence length $L = 100$).

Given the set of N read sequences, the statistical task is to infer the sequence most likely to have generated the observed data. An experiment similar to this model was demonstrated on very short sequences with tunneling current data in (12). Here we extend the approach to the longer sequences expected from a nanopore device.

In our HMM formulation, each output read is modeled as a discrete set of observed states, $\mathbf{x} = \{x_1 \dots x_T\}$, $x_i \in (A, C, G, T)$, a vector of observed bases, and a discrete set of hidden states, $\mathbf{z} = \{z_1 \dots z_T\}$, $z_i \in (1 \dots L)$, the unknown position along the sequence. An HMM is described by three model parameters, the initial state distribution $\pi = p(\mathbf{z}_1)$, the hidden state transition matrix $A = p(\mathbf{z}_t \mid \mathbf{z}_{t-1})$, and an emission distribution $S = p(\mathbf{x}_n \mid \mathbf{z}_n)$. π and A are fixed by the experimental conditions. The elements of A are obtained by numerically integrating Eq. 2 over possible transitions, δ. The inference problem in this model is to estimate the emission distribution, S, which acts as an implicit representation of our sequence,

$$S_{dl} = (1 - \epsilon)p(\mathbf{x}_n = d \mid \mathbf{z}_n = l) + \epsilon/4 \qquad (3)$$

In practice, S is a $4 \times L$ matrix with a multinomial distribution over the possible nucleotides at each position (see Fig. 2). We use the expectation-maximization algorithm to maximize the likelihood, $p(X \mid \theta)$, with respect to the model parameters (14). The joint probability of data and states can be written as a product over the independent output reads, $p(\mathbf{X}, \mathbf{Z} \mid \theta) = \prod_n p(\mathbf{X}_n, \mathbf{Z}_n \mid \theta)$, from which follows that we can perform expectation updates on each read individually before averaging results in the maximization step (see Supporting Material for full model derivation). The resulting shared parameter estimation scheme incorporates all reads while allowing an efficient, parallel calculation.

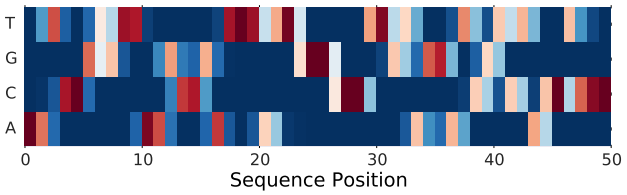

FIGURE 2 An example of the output sequence inference distribution, S. Taking argmax of each columns yields the sequence inference.

After a convergence criterion on the likelihood is satisfied ($\Delta LL < 10^{-5}$), we recover an estimated emission distribution S, which we convert to an estimate for the DNA sequence by taking $\max_{\mathrm{d}} S_{dl}$. The final inference accuracy is measured as the Levenstein distance between the input sequence and inferred sequence, normalized by L. The algorithm has complexity $O(NLT)$, where T is $O(L)$ in the limit of $\beta = 1$. A run of 100 reads of $L = 1$ kb DNA completes in under fifteen minutes. An example of the output of this algorithm showing the relationship between the true sequence and the inferred sequence distribution is shown in Fig. 2 *A*.

We examined the performance of the algorithm over a range of parameter values in order to identify a minimal experimental configuration capable of sequence inference at a given accuracy. First, we consider how to select realistic values for our parameters. $\xi = 0$ corresponds to completely symmetric diffusion and is unlikely to reach the end of the sequence without first exiting from the *cis* side of the nanopore (these reads are discarded in the data generation). $\xi \to \infty$ corresponds to nondiffusive motion and is a trivial case in this model. Lu et al. report an experimental bias term of $\xi = 0.2$, for which they show single-pass accurate sequence recovery is impossible (5). We take this as a starting point for investigation, exploring the range $\xi = 0.2$ to $\xi = 10$ (from $f \approx 2.2$MHz to $f \approx 112$MHz at $\gamma = 12.7$). An appropriate base-call error rate will be device specific and is presently unknown, so we examine the range $\epsilon = 0.0$ to $\epsilon = 0.5$. The number of reads, N, is potentially unlimited, and we examine until convergence.

A sweep across the parameter space is shown in Fig. 3. In each sweep, the error rate was set to $\epsilon = 0.05$ unless otherwise specified. In Fig. 3 *A* we vary the forward bias for a fixed number of reads. The strongest determinant of inference accuracy is the forward bias, controlling how diffusive the motion is. However, the effect of multiple reads is immediately apparent, improving the accuracy markedly until beginning to saturate above $N = 100$. This is expected because each read contributes an independent observation of the input DNA sequence. Accuracy begins to saturate around $\xi = 1$, impressive given the probability of a single base for-

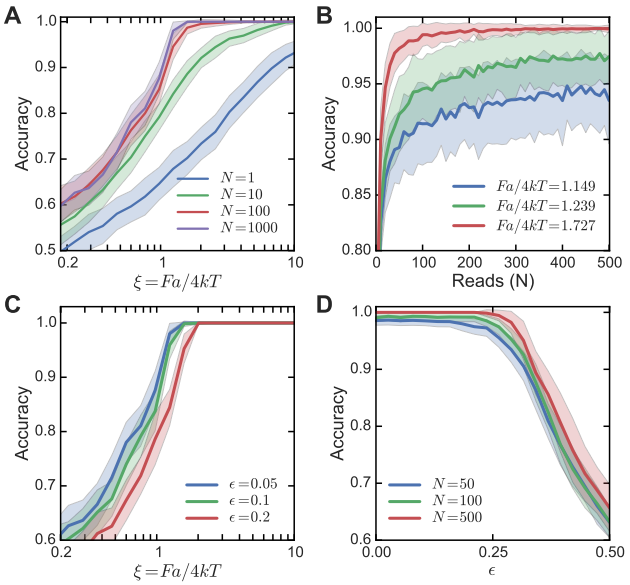

FIGURE 3 Parameter Sweeps. Inference accuracy (measured as 1 - Levenshtein distance/length) is plotted as a function of experimental parameters. $\epsilon = 0.05$ unless specified. (A) Sweep across forward bias, β. (B) Sweep across number of reads, N. (C) Asymptotic performance (large N) performance vs forward bias β. (D) Sweep across error rate, ϵ.

ward transition, $p(\delta = +1)$, is only 0.52. In Fig. 3 *B* we plot the relationship between forward bias and the number of reads in the region of $\xi = 1$. Again, accuracy improves with an increasing number of reads until convergence. Finally, inference accuracy is robust to base-call error rates up to 25%, provided each nucleotide in the sequence is visited sufficient times (Fig. 3 *C* and Fig. 3 *D*). From this data, we conclude a threshold accuracy of 99% is achievable for $\xi > 2$ at an error rate $\epsilon = 0.2$. Following (4) and letting $\gamma = 12.7$ (???), this corresponds to a minimum sampling frequency of $\approx$ 20MHz, well within todays measurement bandwidths. Previous estimates have placed a threshold for accurate sequencing in the GHz sampling range; these results suggest accurate sequencing is possible at sampling frequencies nearly two orders of magnitude lower.

Unidirectional motion of the DNA sequence inside the nanopore has been assumed a necessary prerequisite for accurate nanopore sequencing. We have demonstrated that accurate inference is possible, even in the presence of diffusive motion, with the integration of data under an appropriate statistical model. The strength of the method lies in independently modeling multiple reads of the input sequence and then iteratively combining to yield a joint estimate of the true sequence. More complicated translocation dynamics can be incorporated in the model through modification to the state transition matrix. Under this model, accurate sequence inference is achievable with modest improvements to the experimental constraints of today's nanopore devices ($\xi \approx 0.2$). Designers of nanopore sequencers should emphasize improving sensor bandwidth, as techniques such as the random walk model can account for diffusive motion and maintain high accuracy reads.

SUPPORTING MATERIAL

The full inference model is available at BPJ Online. Source code used to perform the analysis will be available at http://stochseq.github.io.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge helpful discussions with David Blei, Frank Wood, David Pfau, and Peter Sims.

KJE and CHW were supported by NIH grant U54-CA121852 (National Center for Multiscale Analysis of Genomic and Cellular Networks). JWM was supported through the NWO Rubicon Fellowship 680-50-1016. KLS was supported by NIH grant R01-HG006879.

REFERENCES and FOOTNOTES

1. Branton, D., Deamer, D. W., Marziali, A., Bayley, H., Benner, S. A., Butler, T., et al. 2008. The potential and challenges of nanopore sequencing. *Nat. Biotechnol.* 26:1146—1153.
2. Winters-Hilt, S., and M. Akeson. 2004. Nanopore cheminformatics. *DNA and Cell Biology* 23:675–683
3. Venkatesan, B. M., and R. Bashir. 2011. Nanopore sensors for nucleic acid analysis. *Nat. Nanotechnol.* 6:615—624.
4. Larkin J., R. Henley, D. Bell, T. Cohen-Karni, J. Rosenstein, M. Wanunu. 2013. Slow DNA transport through nanopores in Hafnium Oxide membranes. *ACS Nano* 7:10121–10128.
5. Lu, B., F. Albertorio, D. P. Hoogerheide, and J. A. Golovchenko. 2011. Origins and consequences of velocity fluctuations during DNA passage through a nanopore. *Biophys. J.* 101:70—79.
6. Luan, B., H. Peng, S. Polonsky, S. Rossnagel, G. Stolovitzky, and G. Martyna. 2010. Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. *Phys. Rev. Lett.* 104:238103.
7. Olasagasti, F., K. R. Lieberman, S. Benner, G. M. Cherf, J. M. Dahl, D. W. Deamer, and M. Akeson. 2010. Replication of individual DNA molecules under electronic control using a protein nanopore. *Nat. Nanotechnol.* 5:798–806.
8. Cherf, G. M., K. R. Lieberman, H. Rashid, C. E. Lam, K. Karplus, and M. Akeson. 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5Å precision. *Nat. Biotechnol.* 30:344–348.
9. Lubensky, D. K., and D. R. Nelson. 1999. Driven Polymer Translocation Through a Narrow Pore. *Biophys. J.* 77:1824–1838
10. Timp, W., J. Comer, and A. Aksimentiev. 2012. DNA base-calling from a nanopore using a Viterbi algorithm. *Biophys. J.* 102:L37—L39
11. O'Donnell, C. R., H. Wang, and W. B. Dunbar. 2013. Error analysis of idealized nanopore sequencing. *Electrophoresis* In press.
12. Ohshiro, T., K. Matsubara, M. Tsutsui, M. Furuhashi, M. Taniguchi, and T. Kawai. 2012. Single-Molecule electrical random resequencing of DNA and RNA. *Scientific Reports* 2:1–7.
13. Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *P. IEEE.* 77:257—286.
14. Baum, L. E., T. Petrie, G. Soules, and N. Weiss. 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. *Ann. Math. Stat.* 41:164—171.
15. Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likeli-

hood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B. Met.* 39:1—38.