Statistical Inference for Nanopore Sequencing Under a Biased Random Walk Model.

Kevin Emmett,[†*] Jacob Rosenstein,[¶] Jan-Willem van de Meent,[‡] Ken Shepard,[§] and Chris 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 considered a barrier to high accuracy reads. Experimental efforts have sought, with some success, to control DNA translocation to ensure unidirectional motion. The bidirectional motion of DNA inside the pore is a previously unconsidered source of error in alignment algorithms. Here we develop a method of statistical inference that explicitly accounts for this error. We present a statistical analysis demonstrating that highly accurate ($>$99.9%) sequence inference is feasible even under highly diffusive motion, in a biased random walk model. The model uses a Hidden Markov Model (HMM) to analyze multiple stochastic reads, using the Expectation-Maximization (EM) algorithm to infer the most likely sequence. Using this model, we place bounds on the achievable inference accuracy under a range of experimental parameters. We conclude that high accuracy inference is achievable within todays experimental constraints.

Received for publication 1 June 2013 and in final form 1 June 2013.

*Correspondence: kje@phys.columbia.edu.

Rapid advancements 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. Making sense of this data has relied on the development of new algorithms in bioinformatics to process the raw short-read data into usable sequence assemblies. However, the high coverage required for reliable assembly makes *de novo* applications difficult. Additionally, existing methods aggregate over large populations of nominally identical molecules. Applications which require higher molecular resolution, such as rare variant detection, haplotype phasing, and metagenomics often remain out of reach. The development of highly accurate, single molecule, long-read sequencing technologies is critical for sustained progress in these applications (1).

Nanopore sequencing has emerged as a candidate to supercede current generation sequencing and allow for theoretically unlimited read length (2). 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. To date, a significant obstacle of this approach has been overcoming the fast stochastic motion of the individual molecules as they are driven through the pore (3,4). Ideally a device would control the passage of DNA through the pore such that motion was unidirectional and each base had 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 (5–7). Unidirectional motion remains difficult to reliably achieve, leading to a source of error in the read sequence not previously adressed by existing algorithms.

In this letter, we consider a simple physical model of DNA translocation as a one-dimensional biased random walk and analyze how the diffusive molecular motion affects the difficulty achievable read accuracy of nanopore sequencing. To do so, we simulate noisy data where the true sequence is unknown, and develop a statistical technique to estimate the most likely DNA sequence associated with the simulated read signal. Our method is based on hidden Markov models (HMMs), which have recently been used to study multi-base resolution in a nanopore sequencer (8), but have not yet been applied to the problem of diffusive motion inside the pore. This work shows that combining multiple reads from a given input DNA sequence allows accurate estimates of the sequence, both in the presence of highly diffusive molecular motion and at high base-call error rates.

We model motion through the nanopore as one-dimensional diffusion with driving force F and unit base step size a, which results in a random walk with fixed forward bias (9)

$$p = \frac{1}{2} + \frac{Fa}{4k_B T}. \tag{1}$$

Given an input DNA sequence of length L, we generate a single output read by stepping through the sequence with a bias p, at each step making a base-call with error probability e, 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. From this simple physical model, we generate a set of N simulated reads X_n, which each have a unique length $T_n \geqslant L$ that depends on the forward bias. Here N is a notion of sequence coverage in terms of identical molecules sequenced. Coverage 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*. In Fig. 1 *B* a series of random walks at several different forward biases is shown, along with the resulting read length distribution. The expected length of a read is

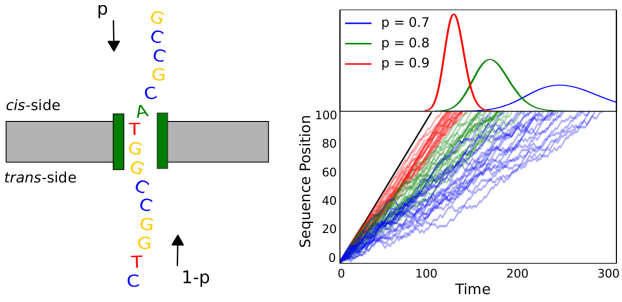

FIGURE 1 A. Schematic representation of an idealized nanopore sequencing device. B. Transit time distribution of a set of random walks for different forward biases (sequence length $L = 100$).

$$\langle T \rangle = \frac{1}{2(p - 0.5)} L. \tag{2}$$

Given this 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 nanopore sequencer data in (10). 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 $A = p(\mathbf{z}_t|\mathbf{z}_{t-1})$, and an emission distribution $S = p(\mathbf{x}_n|\mathbf{z}_n)$. π and A are fixed by the experimental conditions. The inference problem in this model is to estimate the emission distribution, S, which acts as an implicit representation of our sequence,

$$S_{dl} = (1 - e)p(\mathbf{x}_n = d|\mathbf{z}_n = l) + \frac{e}{4} \tag{3}$$

In practice, S is a $4 \times L$ matrix with a multinomial distribution over the the possible nucleotides at each position

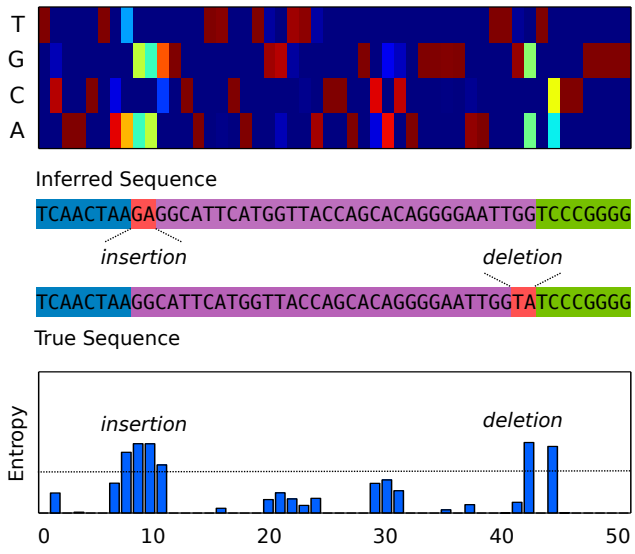

FIGURE 2 The output of the statistical model and its relationship to the true sequence. At top, the output sequence inference distribution, S. Taking the max of each column yields the sequence inference, which we see maps into the true sequence (red) in chunks, with errors in the form of spurious insertions and deletions. The entropy signature, shown at bottom, heuristically identifies these positions.

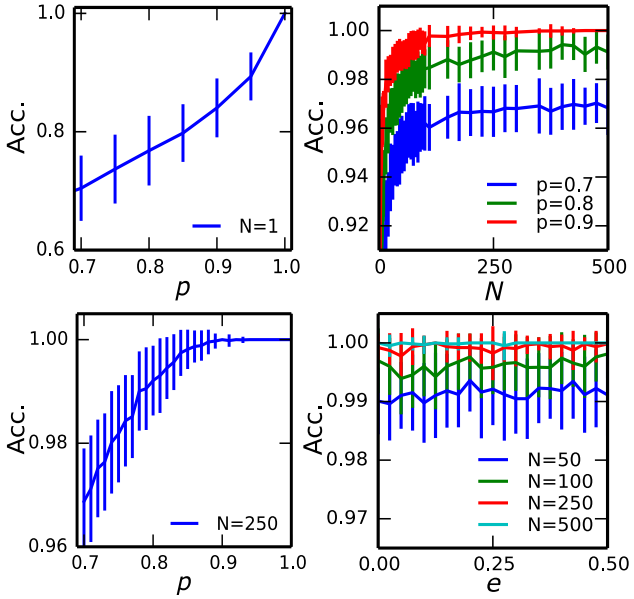

FIGURE 3 Parameter Sweeps. We plot inference accuracy (measured as 1 - Levenshtein distance/length) as a function of experimental parameters. (A) Baseline performance for a single read, $N = 1$. (B) Parameter sweep across number of reads, *N*. (C) Asymptotic performance (large *N*) performance across forward bias *p*. (D) error rate *e* vs *N*.

(see Fig. 2). We maximize the likelihood, $p(X \mid \theta)$, with respect to the model parameters using an implementation of the expectation-maximization algorithm (**?**, 12). The joint probability of data and states can be written as a product over the independent output reads, $p(\mathbf{X}, \mathbf{Z}|\theta) = \prod_n p(\mathbf{X}^n, \mathbf{Z}^n|\theta)$, from which follows that we can perform expectation updates on each read individually before averaging results in the maximization step. The resulting shared parameter estimation scheme incorporates all reads while allowing an efficient, parallel calculation.

After a convergence criterion on the likelihood is satisfied, we recover an estimated emission distribution S, which can be converted 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 $p = 1$, and $O(L^2)$ in the limit of $p = 0.5$. 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*. Additionally, the column-wise entropy of the sequence inference, defined as $H_l = -\sum_d S_{dl} \log S_{dl}$, is shown in Fig. 2 *C*. Inference errors in this model are generally associated with increases in the inference entropy around the error position l, suggesting a possible approach that further analysis might systematically identify and remove the inference errors due to bidirectional motion.

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. $p = 0.5$ 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). $p = 1.0$ corresponds to nondiffusive motion and is a trivial case in this model. Lu et al. report an experimental bias term of $Fa/4k_BT = 0.2$ ($p = 0.7$), for which they claim accurate sequence recovery is near impossible (4). We take this as a starting point for investigation, exploring the range $p = 0.7$ to $p = 1.0$. An appropriate base-call error rate will be device specific and is presently unknown, so we examine the range $e = 0.0$ to $e = 0.5$. The number of reads, N, is potentially unlimited, and we examine until convergence.

A sweep across the parameter space is shown in Fig. 3. For *N*=1, the average inference accuracy goes roughly as ~*p*, a baseline measure of performance (Fig. 3 *A*). In Fig. 3 *B* we plot the relationship between forward bias and the number of reads. As expected, the strongest determinant of inference accuracy is the number of reads, *N*. This is expected because each read contributes an independent observation of the input DNA sequence. We observe an inference error rate that falls rapidly with increasing *N*, even at highly diffusive motion. Generally, the achievable accuracy converges at a fixed *p* even with the addition of more reads. This asymptotic performance is plotted in Fig. 3 *C*. A threshold accuracy of 99% is achievable for *p* as low as 0.8; an accuracy exceeding

99.9% is achievable with $p > 0.88$. Finally, inference accuracy is independent of the base-call error rate, provided the each nucleotide in the sequence is visited sufficient times (Fig. 3 *D*).

Precise control over the motion of the DNA sequence inside the nanopore has previously been assumed a necessary prerequisite for accurate nanopore sequencing. Here we have demonstrated that high accuracy 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 the way in which multiple reads of the input sequence are independently modeled and then iteratively combined to yield a joint estimate of the true sequence. Under this model, reasonably accurate sequence inference is achievable even within the experimental constraints of today's nanopore sequencers ($p \approx 0.7$). Designers of nanopore sequencers may find that there is a tradeoff between sequencing speed and maintaining unidirectional molecular motion. In this scenario, techniques such as the random walk model may allow increased throughput without a loss of accuracy.

SUPPORTING MATERIAL

The full model can be found by visiting BPJ Online.

ACKNOWLEDGEMENTS

The authors acknowledge useful discussions with D. Blei and D. Pfau.

This work was supported by grant XXX (TODO: Chris, need numbers).

REFERENCES and FOOTNOTES

1. Mak, H Craig. 2012. Genome Interpretation and Assembly–Recent Progress and Next Steps. 2012. *Nat. Biotechnol.* 30:1081—1083.
2. 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.
3. Venkatesan, B. M., and R. Bashir. 2011. Nanopore sensors for nucleic acid analysis. *Nat. Nanotechnol.* 6:615—624.
4. 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.
5. 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.
6. 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.
7. 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.
8. Timp, W., J. Comer, and A. Aksimentiev. 2012. DNA Base-Calling From a Nanopore Using a Viterbi Algorithm. *Biophys. J.* 102:L37—L39
9. Berg, H. C. 1993. Random Walks in Biology. Princeton University Press. Princeton, New Jersey.
10. 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.
11. Rabiner, L. R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *P. IEEE.* 77:257—286.
12. 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.
13. Dempster, A. P., N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B. Met.* 39:1—38.