Biophysical Journal: Biophysical Letters

Statistical Inference for Nanopore Sequencing

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, Columbia University, New York, New York; and School of Engineering, Brown University, Providence, Rhode Island

ABSTRACT Nanopore sequencing has been proposed as a possible third generation sequencing platform, however backwards motion of the DNA molecule due to stochastic motion inside the pore is a novel source of error and a barrier to high accuracy reads. We present a theoretical model of DNA translocation through a nanopore as a one-dimensional biased random walk. A Hidden Markov Model is used to infer the input sequence from a set of output sequences. Finally, bounds on inference accuracy are set based on model parameters, demonstrating the accurate sequence inference is feasible even under highly diffusive motion.

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

*Correspondence: kje@phys.columbia.edu.

Current generation sequencing platforms have led to an explosion in available sequence data, rapidly advancing our understanding of the genomic foundations of life. 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, short reads require high coverage for reliable annotation. Even at the highest coverage applications such as *de novo* assembly and metagenomics remain difficult. Additionally, because existing methods do not work at the single-molecule level, some applications, such as haplotype phasing and cancer evolution, remain unreliable. The development of single molecule, long-read sequencing technologies is critical for continued progress in genomics [1].

Nanopore sequencing has emerged as a potential candidate to supercede current generation sequencing and allow for theoretically unlimited read length [2]. A number of strategies for sequencing DNA with nanopores 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 primary obstacle of this approach has been overcoming the fast stochastic motion of the individual molecules as they are driven through the pore [3, 4]. Recent methods have demonstrated an ability to controllably 'ratchet' DNA molecules through a nanopore one base at a time, although experimental results show that motion of the molecule can still occur in both forward and backward directions within a single read [5–7]. Therefore, unidirectional stepwise motion remains difficult to reliably achieve, leading to a novel source of error in the read sequence due to this diffusive motion.

In this letter, we demonstrate a statistical method of combining multiple reads from a given input DNA sequence passing through a nanopore that is able to yield accurate sequence reconstruction, even in the presence of highly diffusive molecular motion. We consider a simple physical model of DNA translocation as a one-dimensional biased random walk, and demonstrate the use a Hidden Markov Model (HMM) to obtain a statistical estimate of the most likely input DNA sequence generating the observed series of reads. HMMs have been previously used to study possible multi-base resolution in a nanopore sequencer [8], but have not been applied to the problem of diffusive motion inside the pore. We then use this model to set bounds on the achievable inference accuracy as a function of experimental parameters. We conclude that multiple parallel reads are sufficient to compensate for both highly diffusive motion and high base-call error rates.

Translocation through the nanopore is modeled as one-dimensional diffusion with driving force F and unit base step size a, which we treat as a biased random walk with fixed forward bias p [9], where p is given by

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

Given an input DNA sequence of length L, we generate a single output read of the sequence by stepping through the sequence with a bias p, at each step making a base call with error probability e. This error rate is strictly due to the base call and is independent of error introduced due to backward motion. Finally, we assume that an appropriate method of making a base call from the raw signal has been implemented. From this simple physical model, we generate a set of N simulated reads, each with length $T^n \geqslant L$ and denoted $X^n_{1:T_n}$. 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 representation of the experimental configuration is shown in Figure 1A. In Figure 1B the distribution of obtained read lengths as a fraction of the input sequence length is shown for several different forward biases, and in Figure 1C an example of the stochastic nature of the transit is shown. The expected length of a read is

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

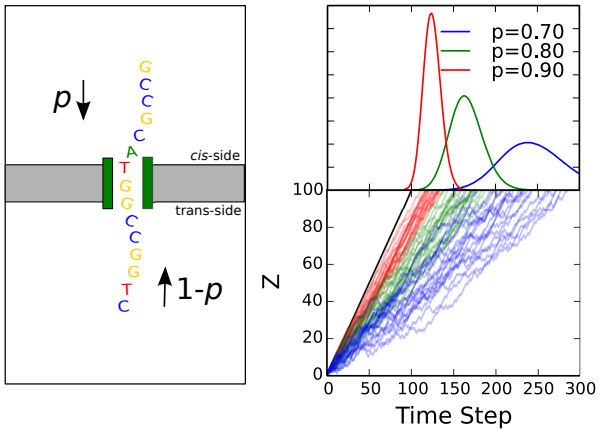

FIGURE 1 A. Idealized example of a nanopore sequencing device. B. Transit time distribution from a set of random walks for different forward biases ($L = 100$).

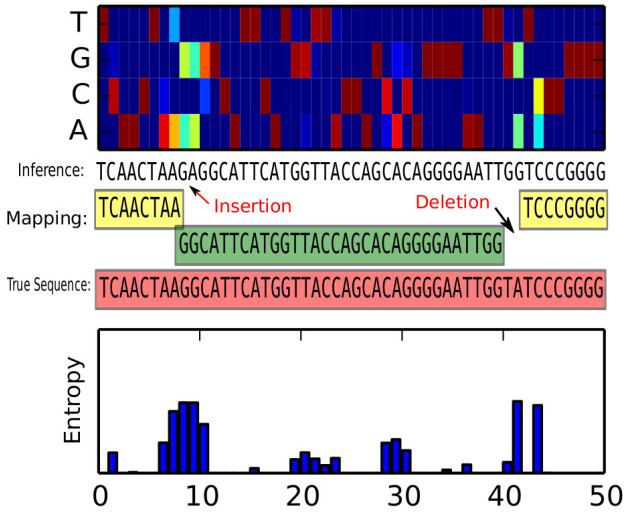

FIGURE 2 Restarts using entropy heuristic. At top, the output sequence inference distribution. Taking $\max$ yields the sequence inference, which we see maps into the true sequence (red) in chunks, errors in the form of spurious insertions and deletions. The entropy signature identifies these positions.

Given this set of N read sequences, the statistical task is to infer the true sequence most likely to have generated the observed data. An experiment similar to this model was been demonstrated on very short sequences with nanopore sequencer data in [10]. Here we extend the approach to longer sequences showing the full power of the approach.

We model the data as a Hidden Markov Model (HMM) [11], where 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. For an HMM, we require 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, and our attention focuses on optimizing $p(\mathbf{x}_n|\mathbf{z}_n)$, which is the explicit representation of our sequence inference. We maximize the complete data log likelihood, $p(\mathbf{X}, \mathbf{Z}|\theta)$, with respect to the model parameters using an implementation of expectation-maximization (EM) [12]. Because the likelihood factorizes over the independent output reads, $p(\mathbf{X}, \mathbf{Z}|\theta) = \prod_n p(\mathbf{X}^n, \mathbf{Z}^n|\theta)$, we can perform expectation updates on each read individually before averaging results in the maximization step. This provides us with a method of jointly using the all the information provided in the reads while still maintaing an efficient, parallel calculation. After a convergence criteria on the complete data log 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 $O(NL^2T)$. An example of the output of this algorithm showing the relationship between the true sequence and the inferred sequence distribution is shown in Figure 2A. Additionally, the column-wise entropy of the sequence inference is shown in Figure 2C. We observe that inference errors are generally associated with increases in the inference entropy, suggesting a possible approach that further analysis might systematically identify and remove inference errors.

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 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 Figure 3. For $N = 1$, the average inference accuracy goes roughly as p, a baseline measure of performance (Figure 3A). In Figure 3B we plot the relationship between forward bias and the number of reads. We observe that, as expected, the strongest determinant of inference accuracy is the number of reads, N. This is expected because each read contributes an inde-

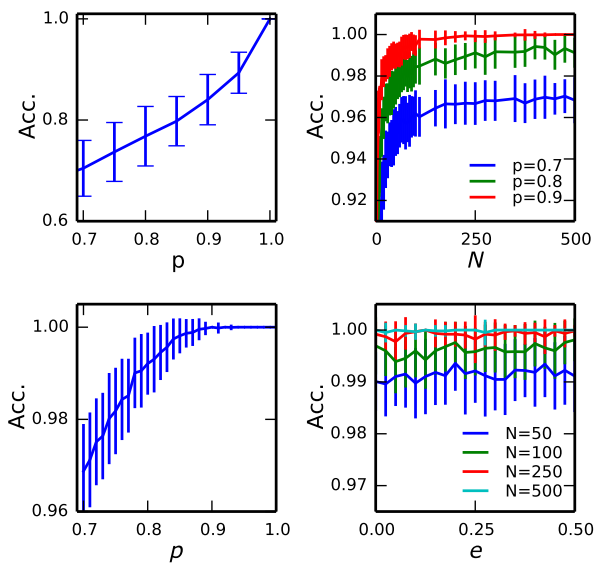

FIGURE 3 Parameter Sweeps. We plot inference accuracy (measured as 1 - edit distance/length) as a function of experimental parameters. (A) Baseline $N = 1$ performance. (B) Parameter sweep across N. (C) Asymptotic performance (large N) performance across p. (D) e vs N.

pendent 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 Figure 3C. A threshold accuracy of 99% is achievable for p as low as 0.8; an accuracy exceeding 99.9% is achievable with $p > 0.88$. High base-call error rates have minimal effect on inference accuracy, provided the number of reads exceeds XXX. (Figure 3D). We emphasize that this nanopore sequencing is designed as a single molecule techinque and as described is not designed for the detection of rare variants.

We have demonstrated a statistical method for inferring the DNA sequence passing through a nanopore sequencer under diffusive motion. 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, 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 motion of the molecules. In this scenario, techniques such as the random walk model may allow increased throughput without a loss of accuracy.

SUPPLEMENTARY MATERIAL

A full derivation of the model and additional supplementary material can be found by visiting BPJ Online at http://www.biophysj.org.

ACKNOWLEDGEMENTS

We acknowledge useful discussions with David Blei and David Pfau. Funding blurb.

REFERENCES and FOOTNOTES

1. Mak, H Craig. 2012. Genome Interpretation and Assembly–Recent Progress and Next Steps. 2012. *Nature Biotechnology.* 30:1081-1083. doi:10.1038/nbt.2425.
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. doi:10.1038/nbt.1495
3. Venkatesan, B. M., and Bashir, R. 2011. Nanopore sensors for nucleic acid analysis. *Nat. Nanotechnology.* 6:615-624. doi:10.1038/nnano.2011.129
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. doi:10.1016/j.bpj.2011.05.034
5. Luan, B., H. Peng, S. Polonsky, S. Rossnagel, G. Stolovitzky, et al. 2010. Base-by-base ratcheting of single stranded DNA through a solid-state nanopore. *Phys. Rev. Lett.* 104:238103. doi:10.1103/PhysRevLett.104.238103
6. Olasagasti, F., K.R. Lieberman, S. Benner, G.M. Cherf, J.M. Dahl, et al. 2010. Replication of individual DNA molecules under electronic control using a protein nanopore. *Nat. Nanotech.* 5:798806. doi:10.1038/nnano.2010.177
7. Cherf, G.M., K.R. Lieberman, H. Rashid, C.E. Lam, K. Karplus, et al. 2012. Automated forward and reverse ratcheting of DNA in a nanopore at 5Å precision. *Nat. Biotechnol.* 30:344–348. doi:10.1038/nbt.2147
8. Timp, Winston, Jeffrey Comer, and Aleksei Aksimentiev. 2012. DNA Base-Calling From a Nanopore Using a Viterbi Algorithm. *Biophysj* 102:L37-L39 doi:10.1016/j.bpj.2012.04.009.
9. Berg, H.C. 1993. Random Walks in Biology. Princeton University Press. Princeton, New Jersey.
10. Ohshiro, Takahito, Kazuki Matsubara, Makusu Tsutsui, Masayuki Furuhashi, Masateru Taniguchi, and Tomoji Kawai. 2012. Single-Molecule Electrical Random Resequencing of DNA and RNA. *Scientific Reports* 2:1–7. doi:10.1038/srep00501.
11. Rabiner, L.R. 1989. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE.* 77:257-286. doi:10.1109/5.18626
12. Baum, Leonard E, Ted Petrie, George Soules, and Norman Weiss. 1970. A Maximization Technique Occurring in the Statistical Analysis of Probabilistic Functions of Markov Chains. *The Annals of Mathematical Statistics* 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. R. Statist. Soc. B.* 39:1-38.