Response to Biophysical Journal Reviews

“Statistical Inference for Nanopore Sequencing with a Biased Random Walk Model”

June 1, 2014

Reviewer #1 wrote:

*The authors apply a sound and sensical approach to the nanopore sequencing problem that they pose. The main issue I have is that the posed problem relies on some very big assumptions which, in my view, are not entirely reasonable. Thus, I am not convinced there is suﬃcient contribution to make this publishable in Biophysical Journal. Here is my detailed critique:*

*1. You assume DNA moves in single nucleotide steps through a nanopore, and that such steps are always resolvable for DNA that moves according to a biased random walk. However, no nanopore sensor works this way. In particular,*

*(a)  This requires unrealistic bandwidths for the nanopore sensor. If DNA moves according to a biased random walk, it does so at single-nucleotide rates that exceed what is possible to measure experimentally. Ref [4] says the minimum bandwidth is 22 MHz to resolve nucleotides at experimental mean velocities they measure, but noise is much too high at this bandwidth for any nanopore sensor to resolve nucleotides. The study shows that reducing voltage force can reduce mean velocity, but DNA motion then becomes dominated even more by thermal agitation and errors induced by backtracking become more pronounced, since DNA is still moving faster than the bandwidth can track (and with nontrivial likelihood of skipping many nucleotides per measurement sample).  The only studies that show DNA motion within measurement bandwidths are the biopore sequencing work (your refs [6],[7]) and in tunnel junction work (ref [12] and others such as by S. Lindsay’s group). In the biopore work, DNA motion is not governed by a biased random walk; instead, Brownian motion is eliminated, and steps are tens of milliseconds per base on average (note base skipping is possible and not uncommon in that work). The tunnel junction measurements are tens of milliseconds per base under certain conditions, but DNA motion is not governed strictly by a biased random walk since there is a force interaction between the tunnel junction and each base.*

*(b)  You assume a tracking mechanism through homopolymer regions would permit detection of single-nucleotide steps. Only tunneling methods promise to supply such a mechanism, though this has not be demonstrated in the literature to my knowledge. For DNA that moves according to your model (in a pore without a force interaction as in [4]), there is no tracking mechanism.*

*How do you resolve these issue? Given your model of DNA motion, what sensor can resolve bases? It seems tunneling methods are closest to what you propose, but this would require a revision to your model to accommodate the force interaction between the base and the sensor. Can your equations be modified to apply in that case? If so, this would be most useful for those methods.*

We appreciate the reviewer’s comments, but we feel that our contribution is independent of these experimental concerns. Our work introduces valuable statistical techniques that complement ongoing improvements to nanopore sensors.

In particular, we believe our paper is the first to demonstrate that it would be possible to achieve high-fidelity sequencing \*without\* fully unidirectional motion of a DNA molecule through a nanopore. In [Lu *et al* Biophys. J. 2011], the calculation of 22MHz as a minimum bandwidth was entirely based on the requirement that no backwards motion could be tolerated. Allowing bidirectional motion will relax this bandwidth requirement considerably. We estimate a required bandwidth of ~2MHz in our model.

As the reviewer notes, single-base motion has been achieved using ratcheting enzymes in combination with nanopores. We feel this contributes to the value and validity of the assumptions in our model. Our introduction of backwards steps in statistical model should be compared to the theoretical contributions of adding base-triplet Viterbi deconvolution as in [Timp *et al* Biophys. J. 2012]. Before multi-base deconvolution proved viable, it was commonly remarked that single-base ionic current resolution was absolutely necessary for nanopore sequencing. Clearly this is not true! Much in the same way, we feel that statistical techniques which deconvolve bidirectional motion of DNA molecules in nanopores will be crucial for improving sequencing fidelity. This implies that sequencing with solid-state nanopores may be more feasible than implied by the reviewer’s comments.

*2. On page 1 you say “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 (57).”*

*It is true that [6,7] demonstrated this, whereas [5] has been proposed but not demonstrated (I suggest you downgrade your reference to [5] as a proposed method).*

Corrected. We have removed the reference to [5] as a demonstrated method.

*3. typo “ distribution of.” (unfinished) just before equation (2).*

Corrected. The time of first passage distribution is the same as the read length distribution.

*4. just after equation (3) you have “ ... over the the ...”*

Corrected.

*5. bottom of page 2 left column, you have figure 2C (there is no 2C).*

Corrected. We have removed the figure of inference entropy.

*6. Below figure 2 you say*

*“Lu et al. report an experimental bias term of Fa/kBT = 0.2 (β = 0.69), for which they claim accurate sequence recovery is impossible (4).”*

*To be fair, they claim single-pass accurate sequence recovery is impossible, which they do justify. So, I would modify the sentence to say:*

*“Lu et al. report an experimental bias term of Fa/kBT = 0.2 (β = 0.69), for which they show single-pass accurate sequence recovery is impossible (4).”*

*You show improvements are possible by rereading, and this complements that work.*

This is an important clarification, and has been corrected in the text.

*7. What ε value is used in Figure 3A-C? Is it zero?*

An error rate ε=0.05 was used in Figure 3. This has been clarified in the figure caption.

*8. Is it true that incorrect base calls are uniformly distributed over the 3 incorrect letters? That seems like a strong assumption that wasn’t explicitly made. This of course would be the reason that, by combining a sufficient number of reads, you can get good accuracy even with “high base-call error rates” as you say in the introduction. If there was bias in the base-call error distribution, which is more likely the case experimentally, I suspect you would have much stronger dependence on ε, with errors growingly with ε obviously.*

A uniform distribution over the 3 incorrect bases has maximum entropy. Biases in the base-call error distribution would have lower entropy and could be calibrated against. This would act to reduce the error rate from substitutions. However, with the modifications to the translocation dynamics we have made, we do indeed observe increasing error with increasing base call error rate. The results appear robust up to approximately 25% base call error rates.

*9. Below figure 3 you say*

*“These results suggest accurate sequencing is possible at driving forces much lower than previously thought.”*

*This is an overstatement. I presume this is in reference to [4], in which case it is fair to say lower forces can be used if there are enough reads, and ONLY IF the sensor can detect each nucleotide with sufficient bandwidth (see point 1 - I don’t think such a sensor exists without adding a sensor-nucleotide force interaction).*

We emphasize that this is a theoretical model that has abstracted from the problem of calling bases in the raw signal. It is a statistical model for dealing with stochastic motion of the DNA in the nanopore. Reference [4] argued that a 50 fold increase in driving force would be necessary for accurate sequencing at the level of only 2 bases (corresponding to a sensor bandwidth of 1.1GHz), however our results suggest that a sensor bandwidth of ~2 MHz will, given enough reads and a reduction in average translocation velocity, be sufficient for accurate sequencing.

*10. Can you say something about the computational requirements in terms of runtime, not just computational complexity? Could this run in real-time, e.g., if each letter is visited for 0.1 microseconds on average? If not, how slow is it? I’m curious. It may be obviously an off-line method, but how long would 100 reads of 1kb take to process? A statement about it would be good.*

The method is indeed an offline method as presently formulated. We have added a statement to the manuscript about runtimes. 100 reads of 1kb takes under five minutes on our local machines with 8 cores, using relatively unoptimized code.

*11. I don’t understand the message behind the last two sentences of the paper:*

*“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.”*

*If DNA goes faster (i.e., with higher force) then throughput goes up (throughput is total amount of sequence per unit time). But how does this relate to accuracy? The work in [4] showed that higher force reduces the influence of Brownian motion and thereby improves errors by making motion more unidirectional and deterministic (though sensor bandwidth requirements become even more unrealistic). On the other hand, your paper suggests you may NOT have to apply these higher forces, but instead reduce the force to experimentally realistic values (the β = 0.7 value established in [4]) and, although Brownian motion creates problems, if you read it enough times your algorithm can tease out the sequence accurately (again, ignoring the sensor bandwidth and single- nucleotide resolution issues). That said, your approach to getting an accurate sequence will make throughput go down. Longer times of DNA in the pore (at lower but realistic forces) and requiring multiple reads both reduce throughput. I don’t see how you can get “increased throughput without a loss of accuracy” under any scenario with your method. Did I misinterpret what you wrote?*

It is true that we are proposing lower throughput relative to a high-force, high-error approach. By increased throughput, we are considering relative to the ratcheting approaches and biopore models previously discussed. We have clarified the wording in the final sentence.

Reviewer #2 wrote:

*Although the manuscript could be of interest, it appears mostly written for a more specialized audience than the Biophysical Journal. The major issue is that some key aspects of nanopore translocaton seem to be lacking in this analysis. For example, the authors only consider read lengths longer or equal to the input DNA length (L). This clearly ignores potential 'misses' due to DNAs that translocate 'too fast' to be detected, which must occur in such a stochastic system. The analysis also assumes that DNA tranlocation is always 'well behaved' (ie, from one end to the other). But, It has been shown that DNA molecules can translocate as u-turns and hairpins, which would result in a significant decrease of sequencing accuracy. Therefore, the numbers reported here do not represent a realistic estimate of error rates and can only be upper limits under an ideal, non-realistic case. For that reason, I cannot recommend this manuscript for publication in the Biophysical journal.*

In response to this criticism, and that expressed by reviewer #1, we have modified our model to explicitly allow for skipped bases. In the new model, we use a Gaussian diffusion kernel to compute the probability of moving between bases at each time step. This allows for nonzero probabilities of skipped bases in either direction. The primary effect of this is to lose sparsity in the transition matrix, slightly increasing the computation time for inference. We have updated our results to reflect this change, and find only a modest increase in the number of reads required for similarly accurate inference. We feel more complex models of DNA translocation, including u-turns and hairpins, are beyond the scope of this model.

*Minor Issues:*

*In equation 2, v is not defined.*

*v* refers to the average translocation velocity. This has been clarified in the text.

*Figure 1B, the areas with the largest number of trajectories do not seem to match the center of the histograms.*

This may have been an artifact of the visualization. In our new model, because motion can both skip and repeat bases, the average read length remains centered around the sequence length, and we have eliminated the subfigure showing read trajectories.

*Convergence criterion in Page 2 (3rd paragraph) is not defined.*

A convergence criteria of ΔLL < 10-5 was used in our analysis. This has been clarified in the text.