

DNA Base-Calling from a Nanopore Using a Viterbi Algorithm

Winston Timp,^{†‡} Jeffrey Comer,[§] and Aleksei Aksimentiev[§]

[†]Department of Biomedical Engineering and [‡]Department of Medicine, The Johns Hopkins University, Baltimore, Maryland; and [§]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, Illinois

Supporting Material

Brownian Dynamics

We performed the Brownian Dynamics (BD) simulations using 3D potential of mean force (PMF) maps to calculate the ion—DNA interactions as described in Comer and Aksimentiev (1). Briefly, the PMF between each ion (K^+ or Cl^-) and each isolated DNA nucleotide (A, T, G, or C) was computed in 3D from a series of all-atom molecular dynamics simulations. Radial ion-ion PMF functions were also computed from all-atom molecular dynamics for each ion pair to model ion-ion forces in the BD simulations.

We created the PMF maps for the simulated BD systems by rigidly transforming the PMF maps for isolated nucleotides and adding them together along with a potential energy function representing the membrane and nanopore. The BD simulation system measured $4.0 \times 4.0 \times 7.2 \text{ nm}^3$. The potential energy of a 1.8 nm-diameter pore in a 3.5 nm membrane was described using eq. 4 from ref. (1). Basepair maps were created by rigidly transforming the nucleotide maps to best match an atomic model of a basepair in a double-stranded DNA. The final models were assembled by adding three basepair maps to the system PMF for each ion. The basepairs were tilted in the pore so that the normal to the plane of the bases was 50 degrees from the pore axis. One basepair was placed in the center of the pore, while the other two were translated -0.65 and 0.65 nm, respectively, along the pore axis. Such a conformation resembles that of a double stranded DNA stretched in a nanopore(2). A total of 128 PMF maps were created for each of the 64 possible DNA sequences and each of the two ions.

We applied a uniform electric field of 180 mV/7.2 nm to the ions in the direction of the pore axis to simulate a 180 mV bias (3). Each system contained 56 K^+ and 50 Cl^- ions, which corresponds to a bulk KCl concentration of 1.3 M, and was electrically neutral. We used constant diffusion coefficients of 1.75 and 1.85 nm^2/ns for K^+ and Cl^- ions, respectively; thus, we did not consider a position-dependent diffusivity of the ions. A 20 fs time step was used to integrate the BD equations of motion. For each DNA sequence, we performed 10 independent BD simulations (each 160 ns in length) with different initial ion positions.

To calculate ion current from the BD simulations, we employed the method described by Aksimentiev and Schulten(3) modified(4) to restrict the calculation to ions within 1.4 nm of the center of the membrane. In this method, the current is computed from ion displacements between consecutive frames of the simulation trajectory. In computing the mean current, the first 4 ns of each simulation was discarded. Therefore, the mean current values plotted in Figure 1A represent an average over 1560 ns.

The quoted uncertainties in the current values were computed by $\Delta = \sigma(\tau)/\sqrt{N}$, where σ is the standard deviation of the current values for all pairs of consecutive frames, $\tau = 0.04 \text{ ns}$ is the time between frames and $N = 38980$ is the number of pairs of consecutive frames. We have found that the standard deviation of the current calculated between two frames depends on the time between frames as $\sigma(\tau) = A(1/\sqrt{\tau})$, where A is a constant. Because the number of pairs of frames $N = T/\tau$, where $T = 1.6 \mu\text{s}$ is the total simulation time, $\Delta = A(1/\sqrt{T})$ and does not depend on τ , the time between frames. Thus, Δ is the appropriate measure for the uncertainty of the currents calculated from the simulations. That is, if one performed another 1.6 μs simulation using a different set of random numbers, there is a high

probability (about 68% since the current values appear normally distributed) that the estimated current from the new simulation would differ by less than Δ .

The statistical uncertainty, Δ , is due solely to fluctuations in the number of ions in pore and thermal motion of these ions. It can always be reduced by increasing the simulation time. Therefore, the range of uncertainties quoted here (~ 5 pA) can be seen as a lower bound for the uncertainty of analogous experimental measurements in which the DNA is held in a fixed conformation for $1.6 \mu\text{s}$. A relevant uncertainty for currents measured in experiments would be the standard deviation of independent current measurements on the same sequence. Due to the fact that the DNA conformation may vary slightly (or significantly, depending on the pore geometry) from measurement to measurement, a much larger lower bound for the uncertainty may exist that cannot be reduced by increasing the time over which the current is sampled.

Simulated Base-Calling

To perform base-calling, we first simulated the current signature using the sequence of interest, either λ DNA or segments of the human genome (hg19; GRCh37). We took each triplet of basepairs and generated a current value using Gaussian random values. The mean and standard deviation of the random values was set using the mean and expected standard deviation observed in the BD simulation of that triplet.

A hidden Markov model was then constructed, using the HiddenMarkov package in R. The emission probabilities for each state were set as Gaussian distributions, with the mean and standard deviation given by the BD simulation for each triplet state. When using the Viterbi algorithm, we set the transition matrix assuming single base steps, i.e. 0.25 for each of the four possibilities described earlier. If we are using only a single current measurement, we set the transition matrix equal for all 64 possibilities (0.015625). Similarly, we set marginal probability for the initial time point to be equal for all 64 possibilities (0.015625). To decode the current signal, we applied the Viterbi algorithm, using the described model, which gave us a series of triplets. We took the center base from each triplet, and reassembled the DNA sequence, then checked the number of bases that were correct.

To vary the noise level, we altered the simulation of the current signature, by applying a multiplier to the standard deviations. By iterating through different values of the multiplier, we measured the effect of SNR on the basecalling efficiency, performing 20 stochastic realizations for each noise level.

To basecall the human genome, we used each complete contig, divided up into 50kb fragments. Each 50kb fragment was called as described above. To determine the length dependence of our basecalling method, we selected a large contig (74 Mb) from human chromosome 1 (30028083-103863906; hg19); we took 30,000 randomly sized fragments from this contig, with replacement. Sequence complexity was calculated using $\sum_i -\left(\frac{f_i}{w}\right) \log_2\left(\frac{f_i}{w}\right)$, where f_i is the triplet frequency for triplet i in a fragment of length w . We calculated this entropy for each fragment of the genome, as a rough estimate of the complexity of each sequence.

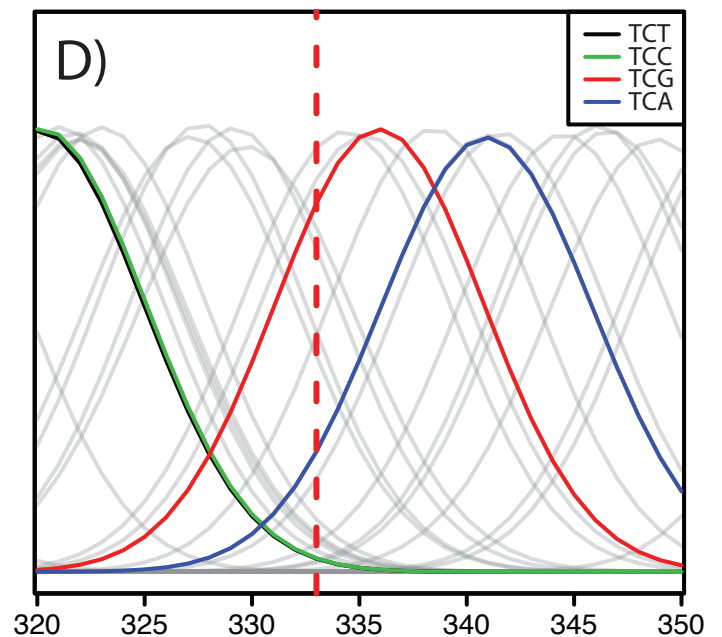
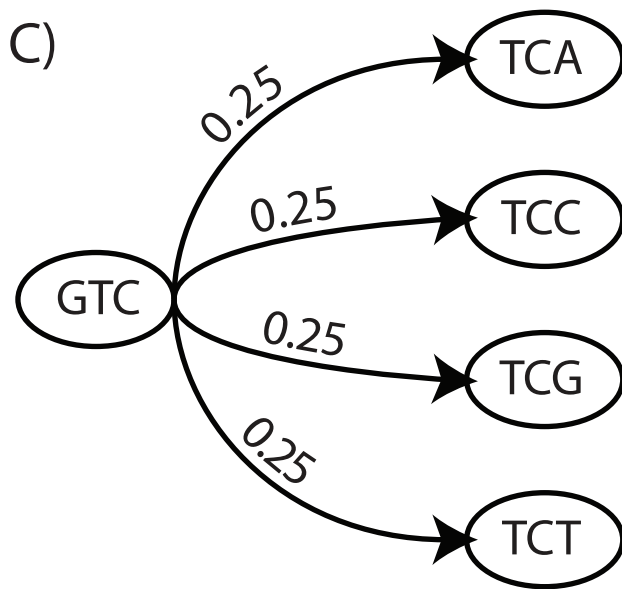
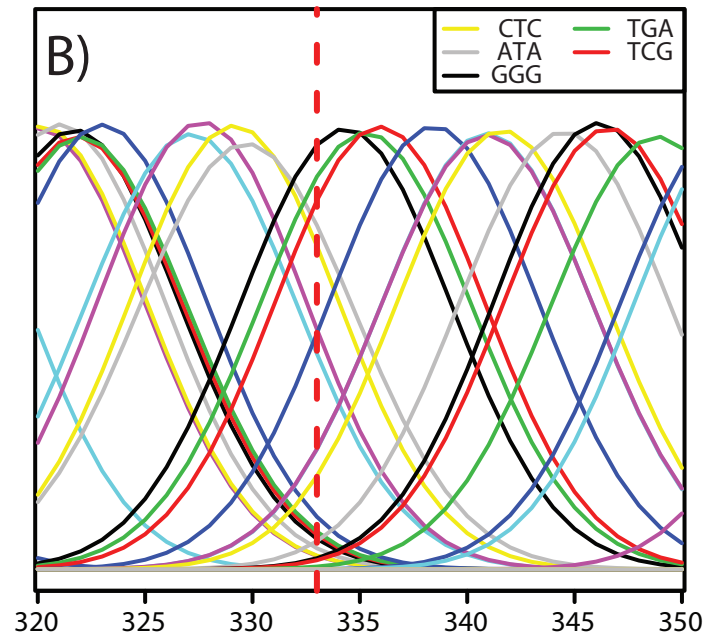
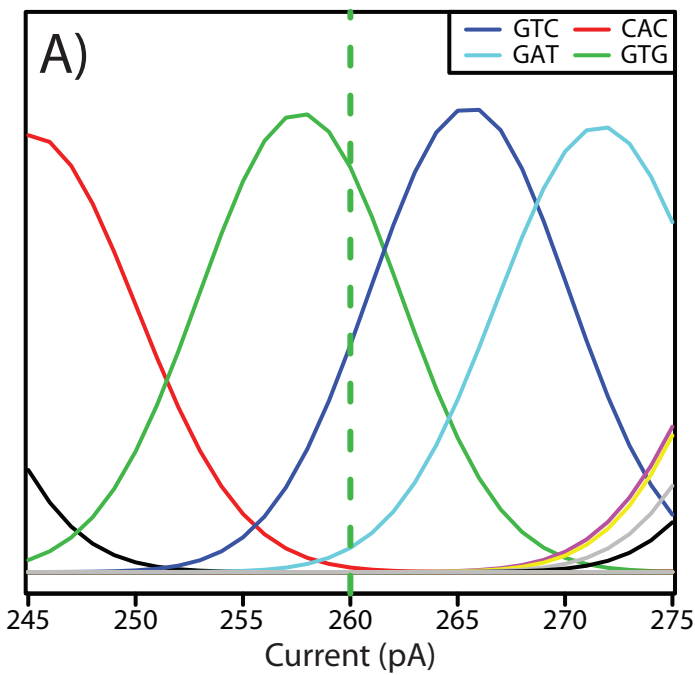
Signal to Noise Ratio Calculation

To calculate signal to noise ratio, we used the ratio of the mean signal level to the standard deviation: $SNR = \mu/\sigma$. To compare to reported experimental results, we used the values reported in Manrao et al.(5). We calculated the standard deviation using the HWHH (half-width half-height) values reported, assuming a normal distribution: $\sigma = (2 \text{ HWHH})/(2 \sqrt{2 \ln 2})$.

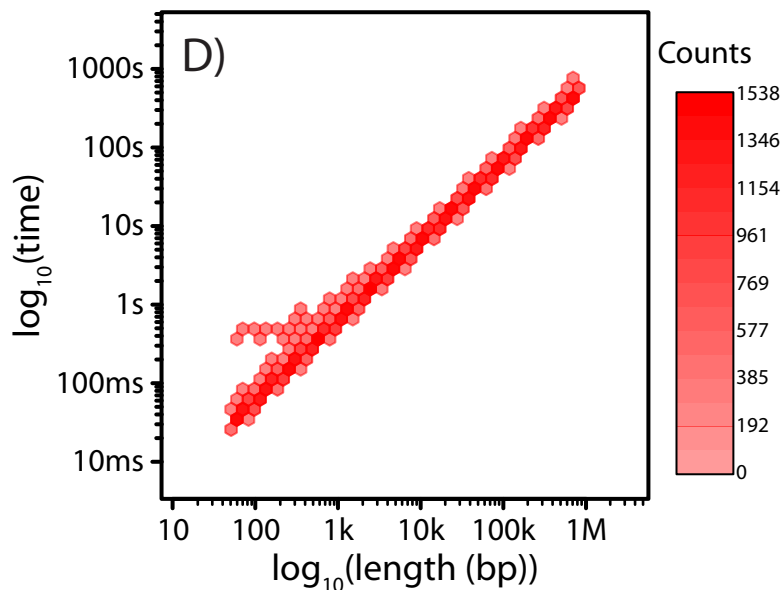
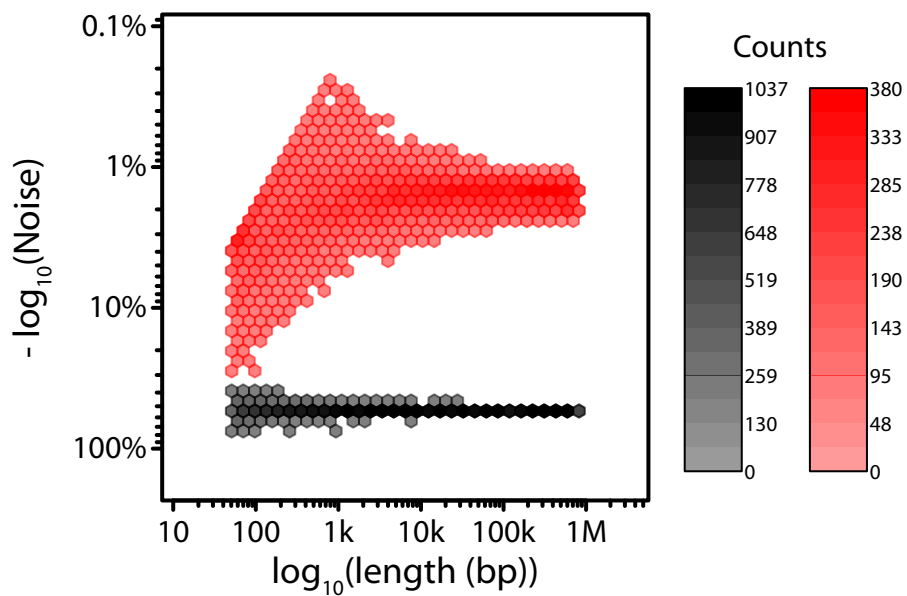
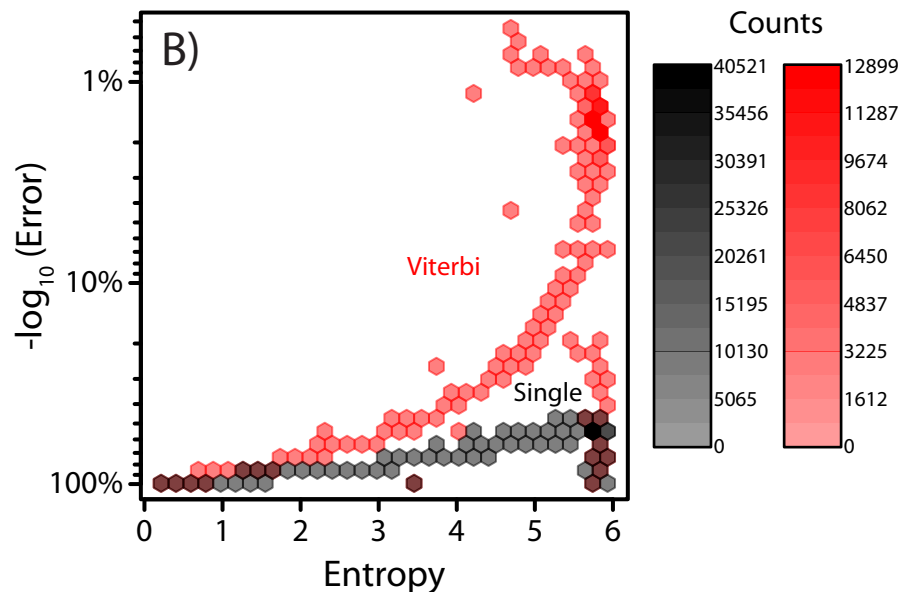
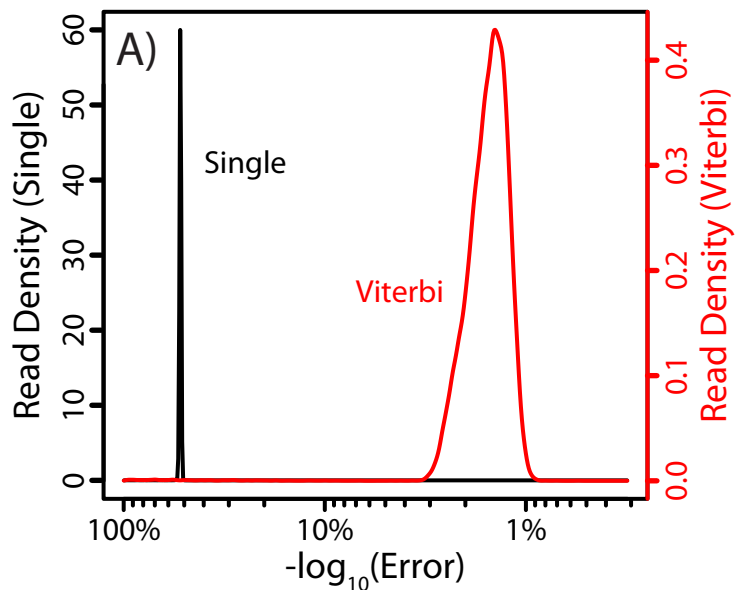
To find the values at which the SNR vs. error curves saturated, we used the derivative of the mean \log_{10} error values (from each noise level), and found the point at which this derivative dropped below 0.01. This was reported as the saturation point.

References:

1. Comer, J. and A. Aksimentiev. 2012. Predicting the DNA Sequence Dependence of Nanopore Ion Current Using Atomic-Resolution Brownian Dynamics. *The Journal of Physical Chemistry C*.
2. Heng, J. B., A. Aksimentiev, C. Ho et al. 2005. Beyond the gene chip. *Bell Labs Technical Journal* 10:5-22.
3. Aksimentiev, A. and K. Schulten. 2005. Imaging alpha-Hemolysin with Molecular Dynamics: Ionic Conductance, Osmotic Permeability, and the Electrostatic Potential Map. *Biophysical journal* 88:3745-3761.
4. Comer, J. R., D. B. Wells, and A. Aksimentiev. 2011. Modeling nanopores for sequencing DNA. *Methods Mol Biol* 749:317-358.
5. Manrao, E. A., I. M. Derrington, M. Pavlenok et al. 2011. Nucleotide Discrimination with DNA Immobilized in the MspA Nanopore. *PLoS One* 6:e25723.



Supplementary Figure 1: Histograms of possible currents for basepairs. **A)** Histogram of current values for different triplets in an easily distinguishable region - a current of 260pA (dotted line) is easily called (correctly) as GTG. **B)** Histogram of current values for a difficult to distinguish region - a current of 333pA is called as GGG, not as TCG. **C)** If we use the information from the previous triplet (in this case GTC) - we can eliminate many of the possible triplet states, as only states which share in common the last two bases of the last triplet (TC) as the first two bases of the new triplet are possible. The probability for each state is given as equal in this case (1/4). **D)** After convolving the probability of the states with the histograms, the calls become much easier. Unlikely states have been reduced to light gray. Note that TCC and TCT are still virtually indistinguishable from each other; if the current was 320pA, a third read would be needed to distinguish the bases.



Supplementary Figure 2: Testing effectiveness of the algorithm using the human genome. **A)** Histogram of log error rate for the Viterbi method (red) versus using only a single measurement (without prior information) (black), from 50 kb fragments of the entire mapped human genome (hg19). Viterbi base-calling has an accuracy of $98.2\% \pm 3.9\%$, compared to $47.4\% \pm 1.9\%$ without using prior information. **B)** Bivariate histogram (hexagonally binned) of Shannon's entropy as a measure of sequence complexity versus log error rate, using the entire mapped human genome (hg19). Error drops dramatically as complexity increases for the Viterbi algorithm. **C)** Bivariate histogram of length versus error rate for random length fragments from (chr1: 30028083-103863906; hg19). Viterbi algorithm is shown in red, single current value in black. **D)** Bivariate histogram of length versus log computational time for base-calling from the same simulation as in **C)**. Log computational time scales as log length.