

Abstract Submitted  
for the MAR13 Meeting of  
The American Physical Society

**Statistical Inference of DNA Translocation using Parallel Expectation Maximization** KEVIN EMMETT, JACOB ROSENSTEIN, DAVID PFAU, AKIVA BAMBERGER, KEN SHEPARD, CHRIS WIGGINS, Columbia University — DNA translocation through a nanopore is an attractive candidate for a next-generation DNA sequencing platform, however the stochastic motion of the molecules within the pore, allowing both forward and backward movement, prevents easy inference of the true sequence from observed data. We model diffusion of an input DNA sequence through a nanopore as a biased random walk with noise, and describe an algorithm for efficient statistical reconstruction of the input sequence, given data consisting of a set of time series traces. The data is modeled as a Hidden Markov Model, and parallel expectation maximization is used to learn the most probable input sequence generating the observed traces. Bounds on inference accuracy are analyzed as a function of model parameters, including forward bias, error rate, and the number of traces. The number of traces is shown to have the strongest influence on algorithm performance, allowing for high inference accuracy even in extremely noisy environments. Incorrectly identified state transitions account for the majority of inference errors, and we introduce entropy-based metaheuristics for identifying and eliminating these errors. Inference is robust, fast, and scales to input sequences on the order of several kilobases.

Kevin Emmett  
Columbia University

Date submitted: 09 Nov 2012

Electronic form version 1.4