

STATISTICS COLLOQUIUM

THURSDAY, APRIL 25, 2013

TALK: 4:15 PM — SCIENCE CENTER RM. 705

RECEPTION: 3:50 PM — SCIENCE CENTER, 7TH FLOOR

“Template-based Estimation of Genome-wide Nucleosome Positioning via Distributed HMC”

Alexander W. Blocker
Department of Statistics
Harvard University

ABSTRACT

The rise of high-throughput scientific experimentation and data collection has introduced new classes of statistical and computational challenges. The technologies driving this data explosion are subject to complex new forms of measurement error, requiring sophisticated statistical approaches. Simultaneously, statistical computing must adapt to larger volumes of data and new computational environments, particularly parallel and distributed settings.

We consider the problem of estimating the genome-wide distribution of nucleosome positions from paired-end sequencing data. We develop a modeling approach based on nonparametric templates to control for the variability along the sequence of read counts associated with nucleosomal DNA due to enzymatic digestion and other sample preparation steps, and we develop a calibrated Bayesian method to detect local concentrations of nucleosome positions. We also introduce a novel set of estimands that provide rich, interpretable summaries of nucleosome positioning. Inference is carried out via a distributed HMC algorithm that can scale linearly with the length of the genome being analyzed. We provide MPI-based Python implementations of the proposed methods, stand-alone and on Amazon EC2, which can provide inferences on an entire *S. cerevisiae* genome in less than 1 hour on EC2. We evaluate the accuracy and reproducibility of the inferences using a factorially designed simulation study and experimental replicates. The template-based approach we present is applicable to single-end sequencing data using alternative sources of fragment length information, such as Bioanalyzer, and to ordered and sequential data more generally. It provides a flexible and scalable alternative to mixture models, hidden Markov models, and Parzen window methods.