Donnelly Centre
for Cellular + Biomolecular Research
UNIVERSITY OF TORONTO

UNIVERSITY OF
TORONTO

Special Seminar

“Deep Learning Frameworks for Regulatory Genomics”

Anshul Kundaje, PhD

Assistant Professor
Department of Genetics
Department of Computer Science
Stanford University

Monday, December 14, 2015 | 10:00 am
Donnelly Centre Red Seminar Room

Abstract:

Assays such as DNase-seq and MNase-seq that profile genome-wide chromatin accessiblity and nucleosome positioning have introduced the possibility of comprehensive identification of regulatory elements (REs) and characterization of their local chromatin architecture. However, computational methods are lacking and experiments are time-consuming, costly and require large amounts of input material, limiting their applicability in rare cell types. We show that multi-task, multi-modal deep convolutional neural networks (CNNs) can be trained using a novel 2D representation of ATAC-seq data that leverages subtle patterns in insert-size distributions to simultaneously predict multiple histone modifications, combinatorial chromatin state and binding sites of a key insulator protein (CTCF) with high accuracy. Models trained on a combination of DNase-seq and MNase-seq data achieve similarly high performance genome-wide and across cell-types supporting a fundamental predictive mapping between local chromatin architecture and chromatin state. We compare the performance of Support Vector Machines, CNNs and stacked CNN-RNNs (recurrent neural etworks) trained on raw DNA sequence to learn various TF binding properties including probabilistic affinity to sequence motifs, positional biases and density of motifs and combinatorial sequence grammars involving co-factor sequence preferences with spacing and order constraints. We show a strong equivalence between biophysical free-energy models of TF binding and CNN based deep learning models. We integrate DNA sequence, DNA shape and chromatin accessibilty to learn predictive models of in-vitro and in-vivo binding for a large compendium of TFs across multiple cell types and tissues. Finally, using novel methods for model exploration, visualization and feature selection, we dissect the heterogeneity of TF binding sites.

Host: Dr. Brendan Frey