Stony Brook University

Biomedical Informatics Grand Rounds

Ramana V Davuluri, PhD

Department of Biomedical Informatics

Stony Brook Cancer Center

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

Wednesday, Sept 30, 2020 3 pm - 4 pm

Abstract: Deciphering the language of non-coding DNA is one of the fundamental problems in genome re-search. Gene regulatory code is highly complex due to the existence of polysemy and distant semantic relationship, which previous informatics methods often fail to capture especially in data-scarce scenarios. To address this challenge, we developed a novel pre-trained bidirectional encoder representation, named DNABERT, to capture global and transferrable understanding of genomic DNA sequences based on up and downstream nucleotide contexts. We compared DNABERT to the most widely used programs for genome-wide regulatory elements prediction and demonstrate its ease of use, accuracy, and efficiency. We show that the single pretrained transformers model can simultaneously achieve state-of-the-art performance on prediction of promoters, splice sites, and transcription factor binding sites, after easy fine-tuning using small task-specific labelled data. Further, DNABERT enables direct visualization of nucleotide-level importance and semantic relationship within input sequences for better interpretability and accurate identification of conserved sequence motifs and functional genetic variant candidates.

In this talk, I will describe pre-training of DNABERT and its application on various sequence prediction tasks, and discuss how DNABERT model can be fined tuned to many other sequence analyses tasks.

Objectives:

1. To understand complex gene regulatory code.
2. To understand the use of deep learning models in deciphering the DNA language.
3. To understand fine-tuning of DNABERT for various sequence prediction tasks,
such as prediction of candidate functional genetic variants.

Continuing Medical Education Credits:

The School of Medicine, State University of New York at Stony Brook, is accredited by the Accreditation Council for Continuing Medical Education to provide continuing medical education for physicians. The School of Medicine, State University of New York at Stony Brook designates this live activity for a maximum of 1.00 AMA PRA Category 1 Credit(s)™. Physicians should only claim the credit commensurate with the extent of their participation in the activity. Disclosure Policy: All those in control of CME content are expected to disclose any relevant financial relationship with a commercial interest (defined as any entity producing, marketing, reselling, or distributing health care goods or services consumed by, or used on, patients) that relates to the content that will be discussed in the educational presentation. All commercial relationships that create a conflict with the planners, speakers, authors' control of content must be resolved before the educational activity occurs. *

Remote Access

Join Zoom Meeting https://stonybrook.zoom.us/j/95617197636?pwd=KytzZ2pVRG9SZGpKZUtpNXJISjNjZz09
Meeting ID: 956 1719 7636 Passcode: 924293
Join by One tap mobile
+16468769923, 95617197636# US (New York)
+13017158592,95617197636# US (Germantown)
Dial by your location
+1 646 876 9923 US (New York) Meeting ID: 956 1719 7636
Find your local number: https://stonybrook.zoom.us/u/abyLdgcObG

Questions? Please call the Biomedical Informatics Department at 631-638-2590.