

Speaker: Yunha Hwang
Affiliation: Harvard University

Talk title: *"Learning to read the genomic language: A machine learning approach to shed light on functional dark matter"*

Abstract: Deciphering the relationship between a gene and its genomic context is fundamental to understanding and engineering biological systems. Machine learning has shown promise in learning latent relationships underlying the sequence-structure-function paradigm from massive protein sequence datasets. However, to date, limited attempts have been made in extending this continuum to include higher order genomic context information. Evolutionary processes dictate the specificity of genomic contexts in which a gene is found across phylogenetic distances, and these emergent genomic patterns can be leveraged to uncover functional relationships between gene products. We trained a genomic language model (gLM) on millions of metagenomic scaffolds to learn the latent functional and regulatory relationships between genes. Our findings illustrate that gLM's unsupervised deep learning of the metagenomic corpus is an effective and promising approach to encode functional semantics and regulatory syntax of genes in their genomic contexts and uncover complex relationships between genes in a genomic region.