MIT Bioinformatics Seminar

Two machine learning models for translating between modalities of single-cell sequencing data and from mass spectra to peptide sequences

Speaker: William Noble, University of Washington

Date: Monday, Nov. 7, 2022

Zoom URL: `https://mit.zoom.us/j/92936688687`

In this talk, I will describe two machine learning translation tasks, one in genomics and one in proteomics. For the first task, translation between modalities of single-cell sequencing data, we propose a semi-supervised translation framework to predict cross-modality profiles. Our Polarbear model is trained using a combination of co-assay data and traditional "single-assay" data. Polarbear uses single-assay and co-assay data to train an autoencoder for each modality and then uses just the co-assay data to train a translator between the embedded representations learned by the autoencoders. With this approach, Polarbear is able to translate between modalities with improved accuracy relative to state-of-the-art translation techniques. As an added benefit of the training procedure, we show that Polarbear also produces a matching of cells across modalities. For the second task, we tackle the longstanding problem of de novo peptide sequencing from tandem mass spectra. We propose a simple yet powerful method for de novo peptide sequencing, Casanovo, that uses a transformer framework to map directly from a sequence of observed peaks (a mass spectrum) to a sequence of amino acids (a peptide). Our experiments show that Casanovo achieves state-of-the-art performance on a benchmark dataset using a standard cross-species evaluation framework which involves testing with spectra with never-before-seen peptide labels. Casanovo not only achieves superior performance but does so at a fraction of the model complexity and inference time required by other methods.