xPore: An AI-Powered App for Bioinformaticians

Tags

https://prod-files-secure.s3.us-west-2.amazonaws.com/dda62111-67f9-4bc2-8d24-91d9290e5fa9/60dc6cdc-3ec9-4165-9ffd-1e9a664253a9/slide_compressed.pdf

Problem Statement

- Nanopore sequencing is a novel technology that can directly sequence RNA molecules
- It measures raw electrical signal data at each nucleotide position along the RNA strand
- This signal data contains information about RNA modifications like m6A methylation
- Goal is to develop a computational method to accurately locate modified positions and quantify modification rates from nanopore data

Data Collection and Preparation

- Nanopore sequencing produces raw signal data in FAST5 format
- Basecalled sequence data is output in FASTQ format
- A reference genome is provided in FASTA format
- These data are preprocessed by aligning to the reference and aggregating into event-level data

- Each event represents the raw signal for a k-mer in the original RNA sequence

Modeling

- A Bayesian multi-sample Gaussian mixture model (GMM) is used
- It models the distribution of electrical signal levels at each genomic position
- Expectation Maximization algorithm used to fit parameters of the GMM
- Each position is modeled as a mixture of 2 Gaussians representing unmodified and modified states
- The model is trained on multiple samples together to improve statistical power

Evaluation

- Model achieves 86% AUC on held-out test set for detecting known m6A modifications
- Provides interpretable modification rate estimates that closely match expected levels
- Performs well on variety of tissue types and cell lines
- Enables discovery of differentially modified positions between conditions

Future Work

- Potential improvements like end-to-end model directly from raw signal data
- Investigate different model architectures like deep autoencoders
- Extend model to detect additional RNA modification types beyond m6A