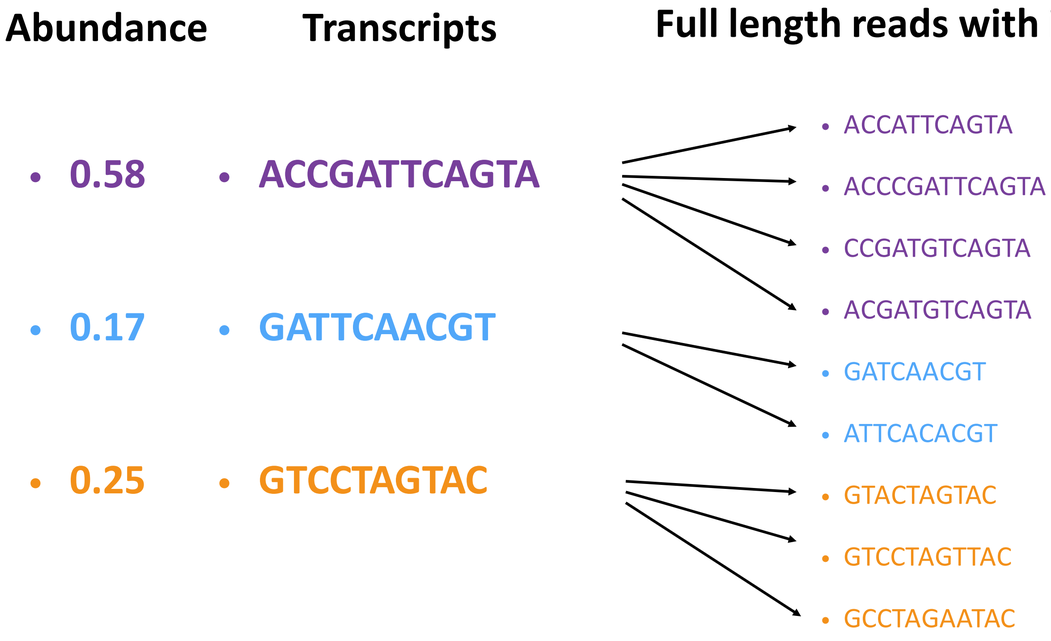
De novo Transcriptome Sequencing via Binary Neural Networks

S. Karen Khatamifard[1], Meisam Razaviyayn[2], Ulya R. Karpuzcu[1]
[1]University of Minnesota, [2]University of Southern California

Abstract

- **Modern Sequencing** platforms: **Billions** of bases per run
- **De novo Transcriptome Sequencing via long reads:** clustering of millions of long RNA sequences, termed reads, based on similarity
- Processing data could take up to **days** even with PacBio's commercial package.
- **Our solution:**
  - Using **neural networks** to find **hashing** functions for obtaining **similarity** → better **scalability** for **parallel** implementation (GPU, FPGA, etc.)
  - Designing a **hardware** accelerator
  - **Binarizing** the network, for a more efficient hardware implementation

Problem Formulation

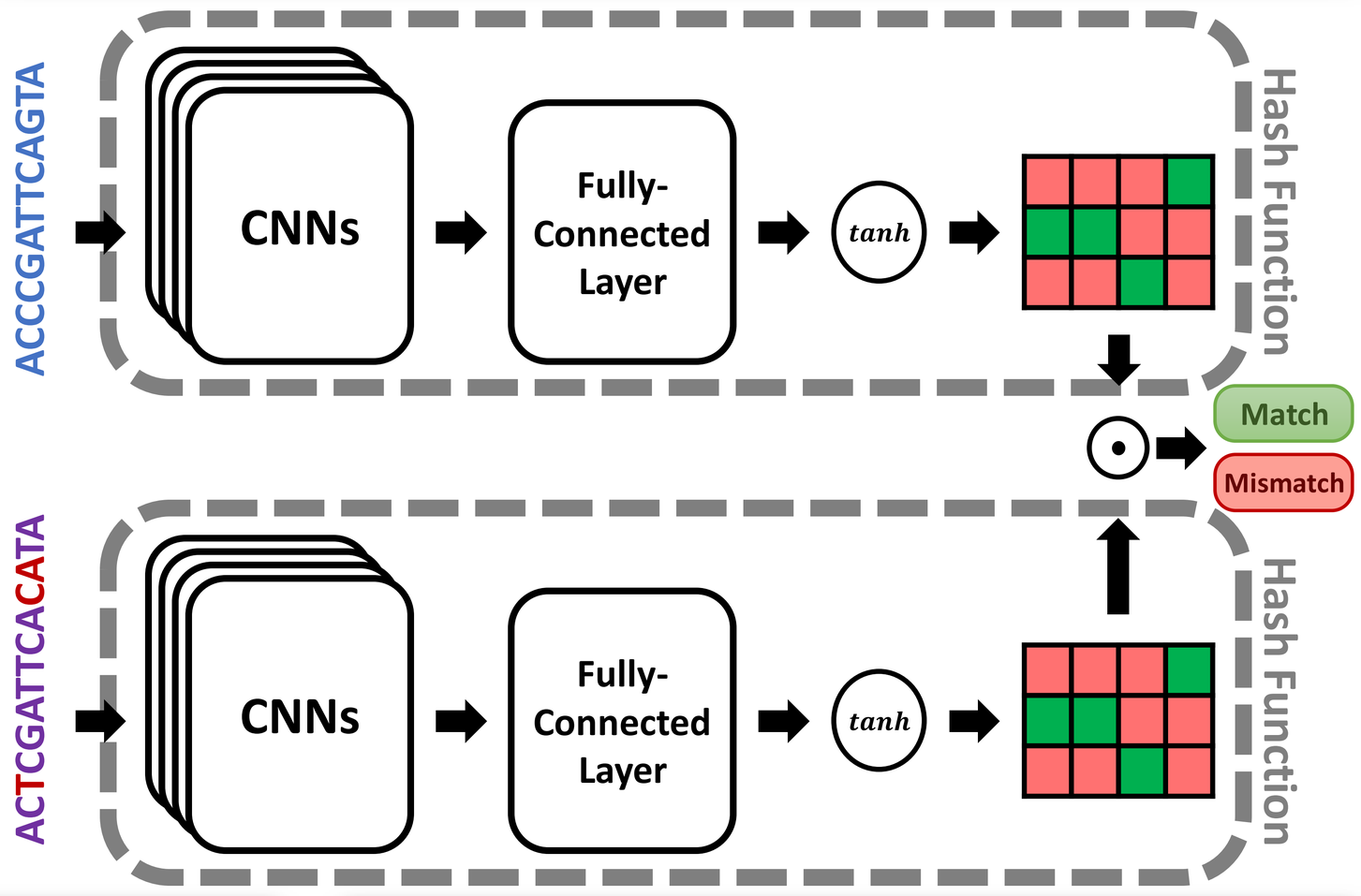


Motivation

- State-of-the-art[1]:
  - Clustering based on obtaining the **similarity graph** between the reads
  - Obtaining the graph:
    - Pairwise similarity computation
    - Similarity kernel → dynamic programing
    - Latency: $\boldsymbol{O(p\,L^2)}$
      - $\boldsymbol{L}$: sequence length
      - $\boldsymbol{p}$: error (mismatch) probability

- Our solution:
  - Using **Neural Networks** to capture similarity
  - Mapping to hardware for acceleration
  - **Binarization** of the network for efficiency
  - Latency: $\boldsymbol{O(H\,log(L))}$
    - $\boldsymbol{H}$: number of neural network layers
    - $\boldsymbol{H} \ll \boldsymbol{L}$

HashNet[2] for Read Similarity



Neural Network Binarization

- Only **convolutional layers** binarized, to maintain accuracy
- **Weight Binarization:**

$$C_{bin}(y, y^*) = \|y - y^*\|_2^2 + \sum_{h=1}^{H} \alpha_h \sum_{w \epsilon W_h} \big((w-1)(w+1)\big)^2$$

- **Activation Function** Binarization [3]:

$$Sign(r), \qquad \frac{\partial Sign}{\partial r} = r\,\mathbf{1}_{|r|\leq 1}$$

- **Deterministic** first convolutional layer
  - 64 masks of length 3
  - Different combinations of {A, C, G, T}

Results

Original	
Accuracy	**# Mult.**
99.1%	**234K**

Binarized	
Accuracy	**# Mult.**
96.3%	**4K**

References

[1] Z. Wang et. al., 2009. RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, *10*(1), pp.57-63.
[2] Z. Cao et. al., 2017. HashNet: Deep Learning to Hash by Continuation. *arXiv preprint arXiv:1702.00758*.
[3] M. Rastegari et. al., 2016, October. Xnor-net: Imagenet classification using binary convolutional neural networks. In *European Conference on Computer Vision* (pp. 525-542). Springer International Publishing.