Qingyang Yin

323-347-1532 yinq@usc.edu linkedin.com/in/qyyin0516 qyyin0516.github.io

Education

University of Southern California 08/2021 – 05/2026 (Expected)
Doctor of Philosophy in Computational Biology and Bioinformatics (GPA: 3.97/4.0) Los Angeles, CA, USA
Courses: Algorithms, Machine learning, Artificial intelligence, Database systems, Computational molecular biology.

Xiamen University 09/2017 – 06/2021
Bachelor of Engineering in Automation (GPA: 3.81/4.0, Ranking: 1/81) Xiamen, China
Courses: Software fundamentals (Data structure), Computer networks, Case analysis for big data, Bioinformatics.

Skills

Programming languages: Python (5 yrs+), R (5 yrs+), MATLAB, C, C++, Java, SQL, Bash
Tools: TensorFlow, Keras, PyTorch, scikit-learn, NumPy, Pandas, MapReduce, Hadoop, Linux, Git
Others: Deep Learning (DNN, CNN, RNN/LSTM, Autoencoder, Transformer, LLM, etc.), Next-Generation Sequencing Toolsets (FastQC, HISAT2, Cell Ranger, DESeq2, SingleCellExperiment, Seurat, etc.), High-Performance Computing

Experience & Selected Projects

University of Southern California, Los Angeles, CA, USA | **Research Assistant** 08/2021 – Present

Cancer driver gene and mutation detection with XAI | *Python* 08/2023 – Present

- Related the mutation status of a gene to its impact on other genes' dependency for cancer.
- Developed a novel **explainable DNN** model with **TensorFlow** to predict the mutation status of frequently mutated genes based on the DepMap cancer dependency data. The AUC can reach 0.96 for some candidate cancer driver genes.
- Designed and implemented a knowledge-primed **autoencoder** to generate driver-like artificial mutations. The earth mover's distance between the mutation presence proportions in artificial mutations and real mutations is 0.004.
- Utilized **activation value differences** and **layer-wise relevance propagation** to find important pathways of driving cancer. 277 out of 343 pathways from the gene-level model exhibit differential expression stochasticity.

Cell type identification with XAI | *Python, R* 08/2021 – 08/2023

- Built an **explainable DNN** model, CellTICS, for cell-type identification from scRNA-seq data using **TensorFlow**.
- Compared the classification performance of CellTICS with other state-of-the-art methods using 6 datasets, demonstrating superior performance with macro F1 scores ranging between 0.9406 and 0.9993.
- Utilized **activation value differences** of neural network neurons to find important pathways of each cell type. On average, uncovered 48 pathways per cell type missed by gene set enrichment analysis.

Machine translation with Transformer | *Python* 05/2023 – 08/2023

- Preprocessed Chinese and English text using dual-tokenization, advanced padding, and truncation.
- Built a Seq2Seq-based translation model to translate Chinese to English by fine-tuning a pre-trained **Transformer** model with **PyTorch**. Trained the model with AdamW optimizer and learning rate scheduling.
- Improved the BLEU score from 42.6 to 51.8 on the test set.

Telomere counting in the genome of Arabidopsis thaliana | *C++* 01/2022 – 05/2022

- Used **DC3 algorithm** to construct the suffix array of input strings extracted from the genome of Arabidopsis thaliana.
- Implemented **Burrows–Wheeler transform** and conducted a query to find the number of exact matches of the telomere sequence in the input. Validated results by comparing with KMP algorithm and checked for linear complexity.

Xiamen University, Xiamen, China | **Undergraduate Researcher** 01/2020 – 06/2021

Feature extraction and classification for scRNA-seq data | *R, Python* 06/2020 – 06/2021

- Employed **stacked autoencoders** with **Keras** for feature extraction and built an **ensemble learning** model to predict cell types and ASD status based on scRNA-seq data. Developed a framework called scIAE.
- Evaluated scIAE against 11 other feature extraction methods through **t-SNE** visualization, **k-means** clustering, and **SVM** classification. scIAE ranked 1st in SVM classification performance.
- Compared scIAE with 15 other classification tools using accuracy, mean F1 score, and median F1 score metrics. scIAE ranked 3rd in accuracy, 4th in mean F1 score, and 4th in median F1 score.

Cell-type-specific predictive models for ASD | *R* 01/2020 – 06/2020

- Applied **recursive feature elimination with cross-validation** to select ASD-related genes from snRNA-seq data.
- Utilized **partial least squares** to predict ASD status of 17 cell types. The average AUC among all cell types is 0.85.

Publications

[1] **Yin, Q.**, & Chen, L. (2024). CellTICS: an explainable neural network for cell-type identification and interpretation based on single-cell RNA-seq data. *Briefings in Bioinformatics*, 25(1), bbad449.

[2] **Yin, Q.**, Wang, Y., Guan, J., & Ji, G. (2022). scIAE: an integrative autoencoder-based ensemble classification framework for single-cell RNA-seq data. *Briefings in Bioinformatics*, 23(1), bbab508.

[3] Guan, J., Wang, Y., Lin, Y., **Yin, Q.**, Zhuang, Y., & Ji, G. (2021). Cell type-specific predictive models perform prioritization of genes and gene sets associated with autism. *Frontiers in Genetics*, 11, 628539.