Haoyun Lei

E-mail: haoyunl@andrew.cmu.edu | Phone: +1(412)969-3798
LinkedIn: linkedin.com/in/haoyunlei/ | Website: https://leovam.github.io/

EDUCATION

Ph.D. in Computational Biology Aug 2016 – May 2021 (expected)
Joint Carnegie Mellon-University of Pittsburgh Ph.D. Program in Computational Biology
Computational Biology Department, School of Computer Science
Carnegie Mellon University (CMU), Pittsburgh, PA, USA
Advisor: Dr. Russell Schwartz
Research Interests: bioinformatics, machine learning, algorithm, discrete optimization, tumor phylogeny, sequencing

B.S. in Biological Science Sep 2008 – June 2012
College of Life Science and Technology
Huazhong University of Science and Technology (HUST), Wuhan, China

SKILLS

Programming Languages: Python (proficient), R (fluent), MATLAB (fluent), Shell (fluent), Java (familiar)

Technical Skills and Tools: Bioinformatics (GATK, SAMtools,bedtools CNVkit etc.), Machine Learning (Sklearn), Deep Learning (PyTorch, TensorFlow), Data Analysis (Numpy, Scipy), Data Visualization (Matplotlib, Seaborn, ggplot2), Combinatorial Optimization (Gurobi, SCIP), Cloud Computing (AWS), Web Development (HTML/CSS/JS)

EXPERIENCE

Ph.D. Project: Integrating multiple data types to infer tumor evolution (Python, R, MATLAB) May 2017 - Present

- Create a mixed membership model for the **Non-negative Matrix Factorization (NMF)** problem
- Develop an efficient **coordinate descent algorithm** to solve the NMF problem in **Python**
- Design a **Mixed Integer Linear Programming Model** with the popular optimization solvers of **Gurobi** and **SCIP**
- Reach **~95% accuracy** with only small set of data and no other existing methods could do this

Benchmarking CNV detection tools (Python, R, Perl) Summer 2020

- Tested and compared public CNV detection tools for calling CNVs in targeted NGS data with a very small panel
- Explored combinations of parameters of tools to increase true positive detection
- Designed algorithms to rescue and recover CNVs with a weaker signal in a very small panel of targets
- Reached over **94%** in sensitivity while kept specificity around **90%**

Converting free-text patient data to ICD codes using natural language processing (Pytorch, TensorFlow) Summer 2020

- Explored language models to annotate and chunk the important information in LabCorp's patient medical text
- Fine-tuned **BERT** model on ICD-10 code classification at chapter and block (first three characters) level
- Designed a **two-step BERT** model to predict multiple ICD-10 codes in LabCorp's patient medical text
- Managed to work on a small dataset and reached **84%** in chapter level clarification

Detection of cancer types and relevant features using deep learning with RNA-seq data (Pytorch) Spring 2020

- Designed and fine-tuned **1D CNN**, **2D CNN** and a **hybrid CNN** models to detect cancer types
- Designed a **Stacked Denoising Autoencoder Classifier** to improve the detections (**~96% accuracy**)
- Applied **embedding** method to find implicit relationships between cancer samples and genes

Footprint Match and Pattern Detection using Machine Learning (Python) Spring 2017

- Classified ~ 10,000 feature matrices with **Neural Network** and **SVM** using **scikit-learn (~95% accuracy)**
- Applied the **Scale-invariant feature transform (SIFT)** algorithm to match of saved and new images
- Extracted the image patterns with **K-Means** and **Gaussian Mixture Model**

Predict Proto Genes using **Logistic Regression, Naïve Bayes Classifier** and **Decision Tree** Spring 2017
Copy Number Extraction from DNA Sequencing Data with **Numpy**, **Scipy** and **Regular Expression** Fall 2016
Model gene regulatory network by combining **Boolean network** and **Ordinary Differential Equation** models Fall 2016

TEACHING EXPERIENCE

Algorithm and Advanced Data Structure Aug 2019 – present

Algorithms: Breadth-first Search, Depth-first Search, Binary Search, Quick Sort, Merge Sort etc.
Data Structure: Linked List, Graph, Tree, Stack, Queue, Heap, ArrayList, Hash Table etc.
Concepts: Recursion, Dynamic Programming, Time and Space Complexity, NP-problem etc.

Laboratory Methods for Computational Biologists Aug 2018 – April 2019

Designed a faster pipeline combining multiple new analysis tools to detect differentially expressed genes in RNA-seq data

BIBLIOGRAPHY

Articles

Tao, Y., **Lei, H.**, Fu, X., Lee, A. V., Ma, J., and Schwartz, R. (2020). Robust and accurate deconvolution of tumor populations uncovers evolutionary mechanisms of breast cancer metastasis. ISMB2020, *Bioinformatics*, *36*, i407-i416,

Lei, H., Lyu, B., Gertz, E., Schäffer, A., Shi, X., Wu, K., Li, G., Xu, L, Hou, Y., Dean, M., and Schwartz, R. (2020). Tumor Copy Number Deconvolution Integrating Bulk and Single-Cell Sequencing Data. RECOMB 2019, *Journal of Computational Biology*, *27(4)* 565-598.

Tao, Y., **Lei, H.**, Lee, A. V., Ma, J., and Schwartz, R. (2020). Neural Network Deconvolution Method for Resolving Pathway-Level Progression of Tumor Clonal Expression Programs with Application to Breast Cancer Brain Metastases. *Frontiers in Physiology*, *11*, 1055.

Lei, H., Gertz, E. M., Schäffer, A. A., Fu, X., Tao, Y., Heselmeyer-Haddad, K., … and Schwartz, R. (2020). Tumor heterogeneity assessed by sequencing and fluorescence in situ hybridization (FISH) data. *bioRxiv*

Tao, Y., **Lei, H.**, Lee, A. V., Ma, J., and Schwartz, R. (2019). Phylogenies derived from matched transcriptome reveal the evolution of cell populations and temporal order of perturbed pathways in breast cancer brain metastases. ISMCO 2019 *(pp. 3-28). Springer, Cham*.

Abstracts & Talks

Lei, H., Gertz, E. M., Schäffer, A. A., Fu, X., Tao, Y., Heselmeyer-Haddad, K., … and Schwartz, R. (2020, July). Tumor heterogeneity assessed by sequencing and fluorescence in situ hybridization (FISH) data. ISMB, virtual

Fu, X., **Lei, H.**, and Schwartz, R. (2020, July). Joint Clustering of single cell sequencing and fluorescence in situ hybridization data to infer tumor copy number phylogenies. ISMB, virtual.

Lei, H., Lyu, B., Gertz, E., Schäffer, A., Shi, X., Wu, K., Li, G., Xu, L, Hou, Y., Dean, M., and Schwartz, R. (2019, May). Tumor Copy Number Deconvolution Integrating Bulk and Single-Cell Sequencing Data. International Conference on Research in Computational Molecular Biology (RECOMB), Washington, DC.

Lei, H., Lyu, B., Gertz, E. M., Schäffer, A. A., & Schwartz, R. (2018, October). Tumor Copy Number Data Deconvolution Integrating Bulk and Single-cell Sequencing Data. In *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, Las Vegas, NV.

Lei, H., Roman, T., Eaton, J., and Schwartz, R. (2018, July). Deconvolution of tumor copy number data using bulk and single-cell sequencing data. Conference on Intelligent System for Molecular Biology (ISMB), Chicago, IL.

Lei, H., Roman, T., Eaton, J., and Schwartz, R. (2018, April). New directions in deconvolving genomics mixtures of copy number variation data. SIAM Conference on Discrete Mathematics, Denver, CO.