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: machine learning, algorithm, discrete optimization, tumor phylogeny

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: Machine Learning, Data Analysis, Algorithm Design, Combinatorial Optimization

System environments: Linux, MacOS, Windows

EXPERIENCE

Ph.D. Projects:

- **Tumor Copy Number Deconvolution Integrating Multiple Types of Genomic Data** 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**
  - Preprocess different data (**PCA**) and visualized the data and results with **seaborn** and **ggplot2** in **Python** and **R**
  - Cluster multiple samples using **Expectation-Maximization** algorithm
  - Reach **~92% accuracy** with small dataset that no other existing methods could do this
  - Published two academic paper on the top conference, and the second one is in preparation

- **Collaboration in Commonwealth Universal Research Enhancement (CURE) project** Dec 2017 - Present
  - Develop models for **large volumes of diverse data** on cancer patients
  - Analyze ~1,000 samples with 100,000 features of DNA, RNA and DNA-RNA data of breast cancer
  - Apply **Medoidshift** pre-clustering and **K-nearest-neighbor-based** reconciliation on the features in **MATLAB**
  - Merge the disconnected subspaces of the features using a **Maximum Likelihood** model
  - Manage to work on **much larger numbers of features** that other highly cited methods could not

Other Projects:

- **Footprint Match and Pattern Detection using Machine Learning** Spring 2017
  - Transformed the real images into feature matrix based on **octonary number system**
  - 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**

- **Prediction of Beneficial Features for Proto Genes** Spring 2017
  - Extracted key features for proto genes using **Logistic Regression, Naïve Bayes Classifier** and **Decision Tree**
  - Tuned the parameters and applied ***k*-fold cross validation** to improve the prediction that previous work didn't find
  - Visualized the rank of features and the analysis of result using **Matplotlib**

- **Copy Number Extraction from DNA Sequencing Data** Fall 2016
  - Analyzed the large-scale genomic data with scientific computing packages such as **Numpy** and **Scipy**
  - Applied **Regular Expression** to match and extract desired information

    - Wrote **Shell** scripts to manipulate files and pass different arguments
    - Obtained a **specific copy number distribution** for a brain cancer (glioblastoma)

- **Dynamic Changes in Gene Regulatory Network** Fall 2016
    - Designed a hybrid model combining **Boolean network** and continuous **Ordinary Differential Equation** models
    - Visualized the regulatory network with **Cytoscape**
    - Estimated more (continuous) states of genes in large-scale network that Boolean or ODE models could not

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, HashMap 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

Peer-reviewed articles

Tao, Y., **Lei, H.**, Lee, A., 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 (Accepted by ISMCO 2019)

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

Abstracts & Talks

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.