Yujie Zhang

Mobile : +1-619-458-8969 Email : yujiezhang@hsph.harvard.edu
linkedin.com/in/yujie-zhang-5a0633152 www.yujiezhang.net

Education

- **Harvard University** Boston, MA
  M.S. in Computational Biology & Quantitative Genetics, Biostats Dept. *Sep. 2021 – May. 2023*
- **University of California, San Diego** San Diego, CA
  B.S. in Bioengineering:bioinformatics, Minor in Mathematics(GPA 4.0), overall GPA 3.80 *Sep. 2017 – May. 2021*

Publications

- A consensus-based and readable extension of Linear Code for Reaction Rules (LiCoRR) *Kellman, B. P.; Zhang, Y.; Logomasini, E.; Meinhardt, E.; Godinez-Macias, K. P.; Chiang, A. W. T.; Sorrentino, J. T.; Liang, C.; ... & Lewis, N. E. Beilstein J. Org. Chem. 2020, 16, 26452662. doi:10.3762/bjoc.16.215* **Co-first author**.
- Correcting for sparsity and interdependence in glycomics by accounting for glycan biosynthesis *Bao, B., Kellman, B. P., Chiang, A. W., Zhang, Y., Sorrentino, J. T., York, A. K., ... & Lewis, N. E. (2021). Nature Communications, 12(1), 1-14. doi:10.1038/s41467-021-25183-5*
- Bacterial modification of the host glycosaminoglycan heparan sulfate modulates SARS-CoV-2 infectivity *Martino, C., Kellman, B. P., Sandoval, D. R., Clausen, T. M., Marotz, C. A., Song, S. J., ... & Armingol, E. (2020). bioRxiv. doi: https://doi.org/10.1101/2020.08.17.238444 Preprint*

Experience

- **Heng Li Lab (Data Science Dept. at Dana Farber Cancer Institute)** Boston, MA
  Research Assistant *November 2021 - Present*
  - **Adjust Hifiasm consistency with Oxford Nanopore Technology (ONT) reads**: Hifiasm is a fast haplotype-resolved de novo assembler for PacBio HiFi reads. It does not perform as ideally on ONT reads as on HiFi reads. I'm working on interpreting the inconsistency and improving its performance on ONT reads.
- **Systems Biology And Cell Engineering Lab (Lewis Lab)** La Jolla, CA
  Undergraduate Researcher *June 2019 - Present*
  - **Linear Code Reaction Rules (LiCoRR) paper**: Linear Code is the most concise and parsable nomenclature for big data analytics of glycans. However, the use of Linear Code by the current field has been inconsistent from each other and from its original setting. In this paper, we are summarizing some accommodations we have seen, together with the original Linear Code implementation rules, to recommend a more consistent version of Linear Code in representing glycosynthesis. We name it Linear Code Reaction Rules (LiCoRR). The paper is published.
  - **Identifying viruses in CHO cells in silico**: Chinese hamster ovary (CHO) cells are widely used cell lines to manufacture protein therapeutics in biopharmaceutical industries. Therefore carrying exogenous viruses is a huge risk for CHO cells. We designed and implemented a computational pipeline to automatically detect and quantify the viruses in CHO cells. One of its many functions is to avoid virus infected cells passing on to biopharmaceutical manufacturing.
  - **Machine Learning - predict glycan motifs**: This project is to predict glycan substructure presence at glycosylation sites given the protein surface. The goal is to apply the program to HIV and COVID-19 data. Through machine learning, we want to figure out the optimal sphere radius for predicting glycans and the best information source for predicting glycans. I'm writing the machine learning algorithms and testing it. (Python + Scipy)
  - **Glyco Analysis Command Line Tool development**: GlyCompare is a program written by PhDs in our lab. It is used to analyze glycans through decomposing them to a minimal set of intermediate substructures. I designed and impemented a command line tool as well as developping additional features for GlyCompare so that it will become a public functional command line tool soon. (Python)
  - **Glycan Database**: Design and implementation of a complex glycan (carbohydrate) database. The lack of interoperability slows the extraction of integrative insight. Our team designed a glycan database which includes all nomenclatures and associated datasets. Therefore, bridging the gaps left by inconsistently identified glycans across datasets and tremendously enriching the information content of the data. (Python + SQL + SPARQL)
- **NanoTools BioScience** La Jolla, CA
  Internship *March 2019 - March 2021*

- **Image Analysis**: Independently Designed and implemented a program based on imageJ using macro language to detect the cardiomyocytes (cardiac muscle cell) contraction change under drug screening. The project decomposed to detection and tracking. The detection is realized by the relative grayscale difference. Then the tracking is realized by Frourier transform based single cell tracking algorithm. The output includes an Excel sheet of contraction data and an animation file denoting the cardiomyocytes contraction directions with green arrows. (ImageJ + Python + groovy) GitHub: https://github.com/yuz682/CardioDT

Professional Activities

- **X Academy** Shanghai, China
  Computational Biology Academic Leader *Aug 2021*
  - **Tutoring and Advising**: Help professors on teaching computional biology to high school and college students in a summer camp. Advising students' capstone project on identifying CpG island on genes by building Hidden Markov Model.
- **Undergraduate Bioinformatics Club(UBIC)** UCSD, CA
  Academic Relations Chair *May 2018 - May 2019*
  - **Chalk Talk Series**: Communicate and invite professors to give speeches on their research fields. Hosted the Chalk Talk
  - **Town Hall Meeting**: Held UBIC annual Town Hall Meeting with faculties to discuss undergraduate bioinformatics curriculum

Award

- **Triton Research and Experiential Learning Scholars (TRELS)** UCSD, CA
  Glycan Database project *Sep. 2019*
- **Triton Research and Experiential Learning Scholars (TRELS)** UCSD, CA
  LiCoRR paper *Jan. 2020*

Skillset

- **Languages**: **Python**, SQL, SPARQL, MATLAB, macro, Java, R, C++, **Linux**, Latex, HTML, Flask
- **Software**: ImageJ, Bash