Toward a Unified Information Framework for Cell Atlas Assembly

Xuegong Zhang
Tsinghua University
Oct. 6, 2021 at MIT Bioinformatics Seminar

Abstract

Profiling the molecular features of all cells with their anatomical and functional attributes is essential for understanding the human body in health and diseases. In recent years, scientists have been enthusiastic in building such atlases of human cells using single-cell omics technologies, led by consortiums and programs like HCA, HuBMAP and HDCA. In the meanwhile, the whole community has been more and more single-cell studies with the rapid development and popularization single-cell RNA-sequencing and other technologies. Tremendous amount of single-cell data are accumulating in the public domain. This suggests the possibility of an alternative approach for building cell atlases by assembling such "shot-gun" data in scattered publications. We have been studying this possibility and the informatics solutions in recent years, and realized that the key challenges in atlas assembly are not unique to scattered data, but also applies data generated by consortium efforts. The task of cell atlas assembly will be of "shot-gun" manner in nature as the spatial and temporal destiny of each cell is not deterministic at the microscopic level. The information complexity and volume are many magnitudes larger than that of the human genome project. We proposed a unified information framework for assembling atlases from data of various sources and built a human Ensemble Cell Atlas (hECA). It includes an infrastructure for storing and retrieving data that are large in both depth and width, and an information graph for unifying and representing multifaceted annotations of cells. We developed an "*in data*" cell sorting scheme that allows extracting cells using logic formula from the "virtual human body" to investigate scientific questions involving multiple organs and cell types. To explore a coordination system that can form a complete representation of the multifaceted nature of cell heterogeneities, we developed a machine-learning method UniCoord for representing multiple discrete, continuous and hierarchical attributes of cells in a unified mathematical space. We hope such a unified representation can enable the quantitative description and investigation of high-order relations across different biological attributes of cells and cell systems. These are ongoing work and the preliminary results have been released as bioRxiv preprints at https://doi.org/10.1101/2021.07.21.453289 and https://doi.org/10.1101/2021.09.09.459281. We welcome suggestions and collaborations to improve the current work.

Brief Bio:

Xuegong Zhang, Ph.D., ISCB Fellow, CAAI Fellow

Prof. Xuegong Zhang earned his BS degree in Industrial Automation in 1989 and Ph.D. degree in Pattern Recognition and Intelligent Systems in 1994, both from Tsinghua University. He joined the faculty of Tsinghua University in 1994, where he is now a Professor of Pattern Recognition and Bioinformatics in the Department of Automation, and an Adjunct Professor of the School of Life Sciences and the School of Medicine. Dr. Zhang studied at Harvard T.H. Chan School of Public Health as a visiting scientist on computational biology in 2001-2002 and 2006, and had been a visiting scholar in the MCB Program at University of Southern California in 2007. He is the Director of the Bioinformatics Division, Beijing National Research Institute for Information Science and Technology (BNRIST). His research interests include machine learning, bioinformatics, intelligent precision health and digit-life twin systems.