Representations of complex variations in multi-scale single cell biological data

Matthew P. Mulè[1, 2]
1. University of Cambridge, Department of Medicine, Cambridgeshire, UK
2. NIH Oxford-Cambridge Scholars Program, National Institutes of Health, Bethesda, MD

Recent technology developments have made routine the large-scale measurement of multiple biomolecules from the same cells at a massive scale. Paired with proliferation of open source analysis software, these new tools have been swiftly adopted across biological disciplines. Experiments using these tools continue to increase in complexity, including measurement of cells from many individuals with different phenotypes or comparison of perturbation responses. These experiments thus capture variations in multiple data modalities covering multiple biological scales; within and between individuals, cell types and across single cells.

These new tools reveal previously unseen detail in biological systems. However, we do not see the cells, their states are first hidden as a matrix of measures which is only brought to light through a lens of statistical abstraction. Strategies to represent data include clustering the cells and projecting clusters onto a two dimensional plane. Some researchers describe these visually striking projections as scientific results in their own right, while others debate whether their role beyond a means of simple visualization, given their uncertain validity in capturing meaningful structure. Other approaches capture differentiation processes, for example ordering cells as a trajectory from a single snapshot and representing cells along a continuum. Descriptions and analysis of these data must balance interpreting data in relation to known cell states, with overinterpretation of structure imposed on the data by statistical algorithms analogous to ET Jaynes' "mind projection fallacy", and thus not representing true ontological states.

Statistical analysis of these modern data have uncertainty at multiple scales, including in whether the computational reconstruction of state maps to a natural cell state, in estimation of variations across multiple human donors, in mapping a inferred cell state to a meaningful emergent property within a cell population in physiological context, and in whether this cell subset

emergent property is correlated with an organism-level phenotype. Here multiple strategies for representing statistical analysis of these data are discussed with emphasis on explicitly representing variations at multiple scales in order to most accurately capture new insights anchored to meaningful cell state ontologies.