CellPress

Principles of Systems Biology, No. 31

This month: selected work from the 2018 RECOMB meeting, organized by Ecole Polytechnique and held last April in Paris.

Single-Cell Data Analysis: Generalizable and Scalable Visualization of Single Cells Using Neural Networks

Hyunghoon Cho and Bonnie Berger, MIT; Jian Peng, UIUC

Algorithmic Advance

Visualization algorithms are fundamental tools for interpreting single-cell data. However, standard methods such as t-stochastic neighbor embedding (t-SNE) are not scalable to datasets with millions of cells, and the resulting visualizations cannot be generalized to analyze new datasets. We introduce net-SNE, a generalizable visualization approach that trains a lightweight neural network to learn a mapping function from high-dimensional single-cell gene expression profiles to a low-dimensional visualization of potentially millions of cells. Our work provides a framework for bootstrapping single-cell analysis from existing datasets (Cho et al., *Cell Systems* 7, this issue, 185–191).

Biological Application

We benchmark net-SNE on 13 different single-cell RNA-seq datasets and show that it achieves visualization quality and clustering accuracy that is comparable to t-SNE while newly allowing previously unseen cells to be mapped onto the same visualization. The mapping function learned by net-SNE can accurately position entire new subtypes of cells and vastly reduce the runtime for visualizing millions of cells.

Our work provides a framework for bootstrapping single-cell analysis from existing data sets.

What's Next?

Inspecting the trained net-SNE models may provide insights into the gene expression patterns underlying t-SNE-like visualizations of single cells. As more single-cell data become available (e.g., https://www.humancellatlas.org/), our work may contribute to learning a low-dimensional "reference" representation of all human cell types to help researchers gain insights about their own datasets.

Statistical Inference of Peroxisome Dynamics

Cyril Galitzine and Olga Vitek, Northeastern University; Pierre M. Jean Beltran and Ileana M. Cristea, Princeton University

Algorithmic Advance

The evolution of the number of peroxisomes in single cells was modeled with a stochastic differential equation with three different rates: the fission generation rate, the *de novo* generation rate, and the degradation rate. The inferred distributions of the rates were obtained from count data simultaneously measured from multiple replicates. A novel and fast parallel inference method based on a particle filter was developed to infer the rates. It directly targets the posterior distribution of the rates while accounting for rate heterogeneity between cells. (Galitzine et al., In Proc. RECOMB 2018, 54–74, https://doi.org/10.1007/978-3-319-89929-9_4).

Biological Application

Our approach can readily determine the peroxisome rates on additional cell types in which peroxisome functions are critical (e.g., neurons). We envision applications to human disease models to better understand peroxisome homeostasis, such as in neurodegenerative disease models. Moreover, application to conditions of cellular stress can give unprecedented detail of the regulation of peroxisomes biogenesis.

We envision applications to human disease models for better understanding peroxisome homeostasis…

What's Next?

The inference procedure can be expanded to other organelles by including additional reactions such as fission. This generalization and the use of multicolor live microscopy to simultaneously visualize several organelles will be explored to achieve a systems view of organelle biophysical properties.

Machine Learning: Feature Exclusion for Digital Tissue Deconvolution

Franziska Görtler, Stefan Solbrig, Tilo Wettig, Peter J. Oefner, Rainer Spang, and Michael Altenbuchinger, University of Regensburg

Algorithmic Advance

Digital tissue deconvolution (DTD) addresses the following computational problem: given a bulk expression profile of a tissue that consists of multiple cell types such as tumor cells, lymphocytes, endothelial cells, or macrophages, what are the abundances of these cells in the tissue? For DTD, we must look at the right set of marker genes—most importantly, genes whose expression differs between tissue and reference must be excluded from analysis. Which are those?

We describe a novel machine learning algorithm that learns the cellular fractions in a tissue while excluding genes for which the linear deconvolution equation does not hold (Görtler et al., In Proc. RECOMB 2018, 75–89, https://doi.org/10.1007/978-3-319-89929-9_5). Our algorithm quantifies large cell fractions as accurately as competing methods and outcompetes them in the detection of rare cell types and in the distinction of similar cell types such as T cell subpopulations.

Biological Application

Our method can be used to quantify cells of a specific type such as tumor cells, B cells, T cells, and macrophages in a tissue using only a bulk expression profile. It can also be used to identify the molecular features that are most informative for such a deconvolution.

For DTD, we must look at the right set of marker genes…

What's Next?

As more single-cell RNA sequencing data become available, our method can learn from these additional data, making DTD more and more reliable.

Metagenomics: Sample Identification via Genome-skimming

Shahab Sarmashghi, Vineet Bafna, and Siavash Mirarab, University of California, San Diego; Kristine Bohmann and M. Thomas P. Gilbert, University of Copenhagen

Algorithmic Advance

Skmer utilizes *genome-skins*, low-coverage genomic sequencing information to rapidly and accurately compute genomic distance between two organisms, and to place an organism in a phylogeny. Skmer is an alignment and map-free approach that uses the k-mer decomposition of reads. It implements a novel method to estimate genome length, sequencing error, and sequence coverage from the k-mer frequency profile, combines them with the Jaccard index between genome-skims to estimate the genomic distances, and uses the distances for phylogenetic inference (Sarmashghi et al., bioRxiv, https://doi.org/10.1101/230409).

Biological Application

The ability to quickly and inexpensively describe the taxonomic diversity in an environment is critical in this era of rapid climate and biodiversity changes. Meta-barcoding via DNA sequencing of taxonomically informative markers has limited phylogenetic resolution. Skmer uses inexpensive low-coverage WGS data and shows great accuracy in estimating genomic distances, finding the exact/closest match to a query sample in a reference set of genome-skims, and phylogenetic reconstruction.

> *Skmer… shows great accuracy in estimating genomic distances, finding the exact/closest match to a query sample in a reference set of genome-skims, and phylogenetic reconstruction.*

What's Next?

We need algorithms to remove contamination from external sources of DNA in genome-skims from biological samples. It remains to be explored if Skmer methodology can be extended to analyze mixed genome-skims of multiple taxa.

Cancer Genomics: UNCOVERing Complementary Functional Cancer Alterations

Rebecca Sarto Basso and Dorit Hochbaum, UC Berkeley; Fabio Vandin, University of Padova

Algorithmic Advance

Identifying groups of cancer alterations with the same functional effect provides insights into the molecular mechanisms of the disease. To identify such groups, we propose a combinatorial formulation for the problem of finding mutually exclusive alterations associated with a functional target. We introduce a tool, UNCOVER, that implements two efficient algorithms to solve the problem and that identifies groups of alterations that may not be identified by analyzing alteration data only (Basso et al., arXiv, arXiv:1803.09721, https://arxiv.org/abs/1803.09721).

Biological Application

On simulated data, UNCOVER finds groups of complementary alterations that are significantly associated with functional targets. On cancer data, UNCOVER finds biologically meaningful groups with stronger association to the target profile than the groups of alterations obtained by the state-of-the-art method and is much faster than the latter. The efficiency of UNCOVER enables the analysis of large datasets from high-throughput functional screens, such as the one from Project Achilles with thousands of target profiles and tens of thousands of alterations. On such a dataset, UNCOVER identifies several statistically significant associations between mutually exclusive groups of alterations and functional profiles, with an enrichment in well-known cancer genes and in known cancer pathways.

> *UNCOVER finds groups of complementary alterations that are significantly associated with functional targets… UNCOVER enables the analysis of large datasets from high-throughput functional screens…*

What's Next?

We anticipate UNCOVER being used to analyze recent high-throughput functional screens. Future work includes the integration of additional data, such as protein-protein interaction networks.

Cancer Genomics: ModulOmics: Integrating Multi-omics Data to Identify Cancer Driver Modules

Dana Silverbush, Tel Aviv University and Simona Cristea, Dana Farber Institute and Harvard (these authors contributed equally); Gali Yanovich, Tamar Geiger, Tel Aviv University; Niko Beerenwinkel, ETH Zurich and Roded Sharan, Tel Aviv University (these authors contributed equally)

Algorithmic Advance

The identification of molecular pathways driving cancer progression is a fundamental problem in tumorigenesis. Its solution can substantially foster our understanding of cancer mechanisms and inform the development of targeted therapies. Current approaches to address this problem use primarily somatic mutations, not fully exploiting additional layers of molecular information. Here, we describe ModulOmics, a method for *de novo* identification of cancer driver pathways, or modules, by integrating multiple data types into a single probabilistic model. ModulOmics simultaneously optimizes the module scores derived from all the data types using a two-step optimization procedure that combines integer linear programming with stochastic search (Silverbush et al., bioRxiv, https://doi.org/10.1101/288399).

Biological Application

ModulOmics identifies highly functionally connected gene modules enriched with cancer driver genes, outperforming state-of-the-art methods. The inferred modules recapitulate known molecular mechanisms and suggest novel subtype-specific functionalities. These findings are supported by an independent patient cohort, as well as independent proteomic and phosphoproteomic datasets.

> *ModulOmics [is] a method for de novo identification of cancer driver pathways… by integrating multiple data types into a single probabilistic model.*

What's Next?

ModulOmics is a flexible framework, allowing the addition of new omics layers to infer cancer driver modules. It can be further extended to integrate multiple datasets in other contexts, such as the inference of protein complexes.

Cell Systems Call

METAGENOMICS: Reconstruction of Microbial Strains Using Representative Reference Genomes

Zhemin Zhou, Nina Luhmann, Nabil-Fareed Alikhan, and Mark Achtman, University of Warwick

Algorithmic Advance

Current reference-based methods yield inaccurate estimates of the species represented in metagenomic sequences due to insufficient sensitivity or multiple false positives. Both are especially problematic for the accurate identification of low-abundance microbial species, e.g., screening for ancient bacterial pathogens in skeletal remains. We present SPARSE, a new method which improves taxonomic assignments of metagenomic reads (Zhou et al., In Proc. RECOMB 2018, 225–240, https://doi.org/10.1007/978-3-319-89929-9_15). SPARSE replaces existing biased reference databases by grouping genomes into hierarchical clusters based on average nucleotide identity (ANI). Reads are assigned to these clusters using a probabilistic model, which further reduces false-positive assignments by penalizing non-specific mappings of reads from unknown sources.

Biological Application

SPARSE yielded greater precision at species-level classification than multiple other methods in multiple simulation studies. For example, SPARSE successfully differentiated multiple co-existing *E. coli* strains in one sample. SPARSE could identify ancient pathogens in archaeological metagenomes with only $\leq$0.02% abundance, while other methods either missed those pathogens or also reported other, non-existent species.

> *SPARSE could identify ancient pathogens in archaeological metagenomes with only $\leq$0.02% abundance…*

What's Next?

SPARSE will be integrated into EnteroBase, an open-access database of 250,000 assembled bacterial genomes from several important pathogens. This will support the reconstruction of the relationships between ancient microbial genotypes within the context of global phylogeography of extant bacteria and support analyses dedicated for reconstructing evolutionary history.