Integrating external information into a graph community detection algorithm to achieve high-quality communities

Ariel Berardino[1], Natalí B. Rasetto[1], Alejandro Schinder[1], and Ariel Chernomoretz[1,2]

[1]Fundación Instituto Leloir
[2]Physics Department, FCEN, University of Buenos Aires/IFIBA (CONICET)

November 8, 2022

Keywords — clustering analysis, community detection algorithm on weigthed graph, continuous process, single-cell/single-nuclei RNAseq technique

1 Introduction

In the field of single cell transcriptomics cells are explored analyzing the density distribution heterogeneity of ensembles of single-cell transcriptional profiles. In this context it is extremely important to identify high quality communities that would led to the discovery of new cell types or cell stages. A general approach to deal with this pattern recognition problem in such a high dimensional space is to focus on a low-dimensional manifold approximation captured by a mutual K-nearest neighbors (MKNN) graph. There are many unsupervised community detection algorithms in graphs that seek to group data-sets according to different figure of merit. However, the community recognition task is an ill-posed problem and different algorithms typically produce different partitions of the data. In this work we address this issue and introduced scBioMerging: a method that integrate external information to identify robust and biologically relevant communities in single-cell transcriptional landscapes.

2 Methods

We aimed to get a biologically meaningful similarity measure between assayed cells. We started from a gene expression matrix obtained in a single-cell RNAseq experiment and constructed a mutual k-mutual nearest neighbor graph (**graphX**) based on the correlations between cells in the Principal Component space. Then, we calculated the standardized transcriptional profile of each cell (Z_i) in graphX. At the same time, we identified over-represented Gene-Ontology Biological Processes (i.e. **external information**) and, for each cell, computed a biological enrichment profile that we used to embed the assayed cells in a kind of biological space. An MKNN graph was then constructed based on the correlation between cell enrichment profiles (**graphBP**). Finally, we computed a **biological process similarity matrix** using a topological measure of similarity from graphBP and used it to weighed the graphX edges. In this way a scalar field that captured the biological-similarity between linked pairs of nodes was incorporated into graphX.

The idea was then to used an heuristic similar to the one implemented in the Louvain algorithm. We started with a high resolution partition of nodes (obtained by applying a k-mean community detection algorithm in PCA space) and we considered the corresponding **community graph** (each node representing a cluster from graphX nodes). Accumulated biological similarity was considered to weigh self-loops and inter-community links in order to look for partitions that maximize the modularity by merging clusters that were biologically similar and produce a new enhanced partition (P_i). This process was repeated about 10 times for different k-means initializations. Finally, a partition (P_f) was generated by applying hierarchical clustering on the adjacency matrix calculated with a voting method that weighs the matrices of the different partitions P_i (**WEAC**)(Dong Huang, 2015).

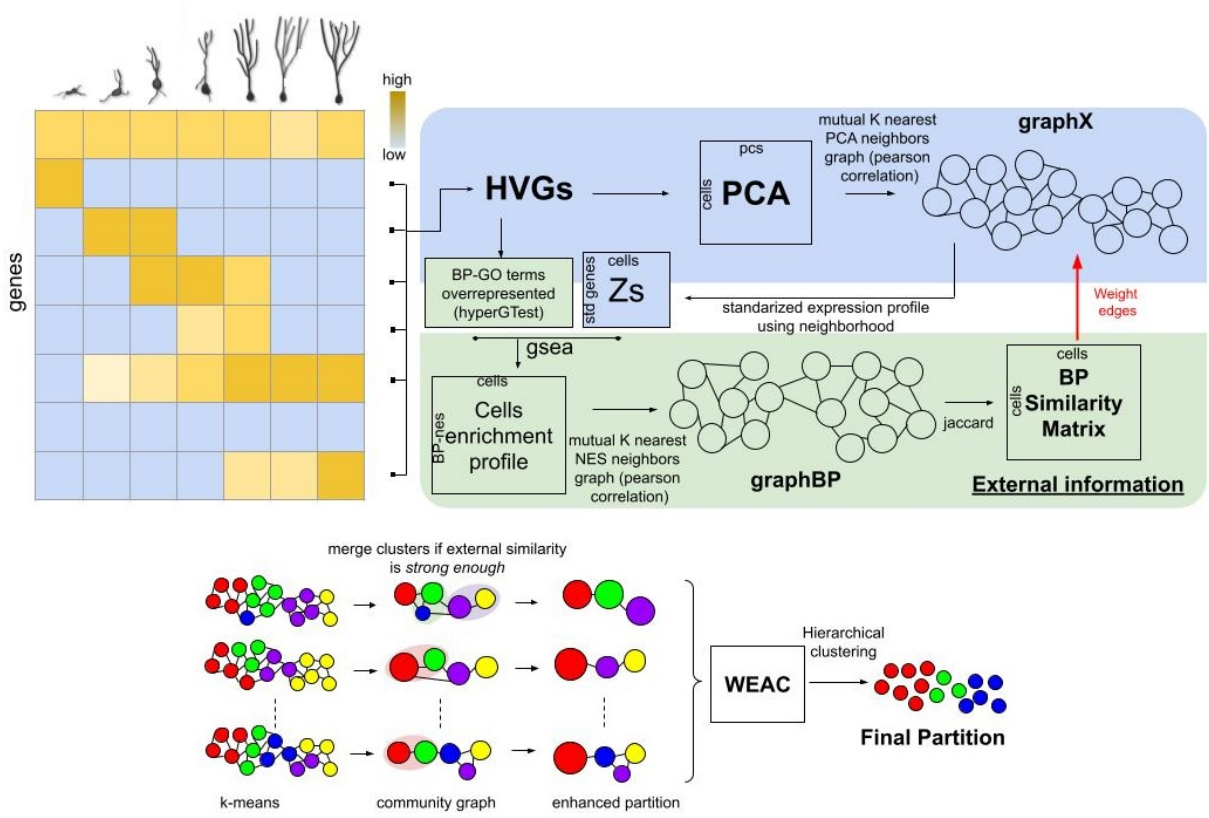

Figure 1: Scheme of scBioMerging's pipeline. We start from top left corner with the gene expression matrix, then we use the highly variable genes (HVGs) to compute de Principal Component Analysis Matrix. Using pearson correlation on the cells embedded in PCA space we built a mutual k-mutual nearest neighbor (MKNN) graph (graphX) on this expression space. We used this graph to calculate the standardized transcriptional profile of each cell in Zs matrix. We also used the HVGs to identify over-represented Gene-Ontology Biological Processes and then we computed a biological enrichment profile to embed the assayed cells in a kind of biological space. We used the same approach as in the expression space to construct a MKNN graph (graphBP) and then calculated a similarity matrix of biological processes for the cells. This matrix served as external information to be inyected in edges of graphX and capture the biological-similarity between linked pairs of nodes. Accumulated biological similarity was considered to weigh self-loops and inter-community links in order to look for partitions that maximize the modularity by merging clusters that were biologically similar and produce a new enhanced partition. This process was repeated about 10 times for different k-means initializations. Finally, a final partition was generated by applying hierarchical clustering on the adjacency matrix calculated with a voting method that weighs the matrices of the different partitions.

3 Results

Using scBioMerging on single-cell and single-nuclei RNAseq developmental datasets, we found clusters that were remarkably similar to those annotated by the authors of published papers (a.k.a "ground truth"). These clusters served as a solid starting point for: the identification of meaningful marker genes or the analysis of differential expression patterns between putative cell types or developmental stages.

The clusters provided by our method served as a solid starting point for: the identification of meaningful marker genes or the analysis of differential expression patterns between putative cell types or developmental stages.

4 Conclusions

We propose a novel and robust method that uses external information to generate a well defined partition on a continuous process. In particular, it can be used to identify biologically relevant cellular stages in a developmental dataset produced by single-cell/single-nuclei RNAseq techniques. We hope that it will help researchers in the analysis of this type of datasets and that they can find important cell stages that have a fundamental role in their developmental study.

References

C.-D. W. Dong Huang, Jian-Huang Lai, Combining multiple clusterings via crowd agreement estimation and multi-granularity link analysis, Neurocomputing 170 (2015) 240–250.