# Clustering in Single-Cell Data

Kelly Jones[1,*] and Andrew Howe[1]

[1]Department of Computer Science, Columbia University

*To whom correspondence should be addressed.

**Abstract**

**Motivation:** The goal of this analysis was to determine whether Affinity Propagation (AP) clustering can accurately cluster single-cell data by comparing it with other clustering methods, such as Louvain and K-means clustering, and employing strategies to address data sparsity.

**Results:** AP clustering did not outperform the Louvain method in terms of Normalized Mutual Information (NMI), but our results show that the Jaccard similarity index with $k$ = 280 transformed with tangent function was the best parametrization for AP clustering. There were 14 clusters in the true assignments set. The Louvain method converged on 19 clusters (NMI = 7.54), and AP using Jaccard similarity converged on 14 clusters (NMI = 0.695).

**Contact:** kaj2165@columbia.edu; arh2207@columbia.edu

**Supplementary information:**

Submission link: https://github.com/arh2207/CompGen_Final_Submit

R package link: https://github.com/arh2207/APJaccard

## 1 Introduction

When making computational inferences on single-cell RNA-seq data, defining communities with similar gene expression in an unsupervised fashion is often necessary for downstream analyses. For this reason, clustering in the gene-expression space is almost always performed at the beginning of single-cell analyses. The Louvain method is the most popular clustering algorithm. The Louvain method detects communities on a nearest-neighbors graph by optimizing a modularity score. SingleR is another popular algorithm which assigns cell-type labels by correlating scRNA-seq data to bulk expression data. Community detection is a highly customizable step, as the resolution of clusters can group large populations of similar cells or distinguish fine-grain subpopulations. Unfortunately, algorithms may poorly separate communities and hence group together cells which were originally distinct in tissue samples. Approaches that cluster from similarity matrices include K-medoids or the K-centers algorithm, which suffer from local minima (Murphy, 2012). Further popular alternative methods of clustering single-cell data include Partition Around Medoids (PAM clustering), which is computationally slow in high dimensions, and Multi-Way K-Means. Affinity Propagation (AP) clustering is a message passing algorithm, which must choose a data point to serve as an exemplar, or cluster center, to define each community (Frey & Dueck, 2007).

Several results reported by Frey and Dueck show that the method significantly outperforms K-medoids. For a dense graph representation of the **S**, AP Clustering may converge in $O(n^2)$ time; but, with sparse similarity matrices, it only takes $O(E)$ time, where $E$ is the number of edges, or non-zero entries, in the similarity matrix **S** (Murphy, 2012). The algorithm therefore may be favorable computationally if we can assign a similarity of 0 to cells that are unlikely to be from the same community. In this experiment, we use the Jaccard similarity index to address data sparsity in single-cell communities. We also explore several parametrizations of the AP clustering algorithm by adjusting the value on the diagonal of **S** and by constructing **S** with different similarity metrics, including negative distances and correlation coefficients. We also transform the Jaccard similarity matrix to differentially weight highly similar or dissimilar data points. We evaluate the quality of clustering solutions using the Normalized Mutual Information (NMI) with respect to an experimentally determined set of cluster assignments (Kozakura et al., 2017). Utilities which allow for further experimentation and generalization of our analysis on arbitrary datasets are implemented as a package in R, available on Github.

## 2 Methods

Data used in this experiment include 10000 individual pancreatic cells from four human donors, collected by Baron et. al. Cells were assigned to different clusters validated by marker analysis and immunohistochemistry stains (Baron et al., 2016). Clusters with abundant populations include β-cells ($n$ = 2407), α-cells ($n$ = 2295), and Ductal cells ($n$ = 958), though 12 other distinct populations were included. Prior to further quality control in this experiment, the original dataset was filtered to remove cells which lack many features, and during data integration, pseudogenes, mitochondrial and ribosomal genes, were removed.

AP clustering is a message-passing algorithm which takes in a matrix of similarities **S** where entry $\mathbf{S}_{i,k}$ is the similarity score between points $i$ and $k$. The algorithm passes messages $a(i,k)$, or the 'availability', and $r(i,k)$, or the 'responsibility', between all pairs of points $i$ and $k$ on each iteration. At the end of the process, the point $i$ which maximizes the following expression is selected for each point $k$ (Frey et al., 2007)

$$\max(a(i,k) + r(i,k)) \qquad (1).$$

If the value of $i$ that maximizes this expression is equal to $k$, then point $k$ is an exemplar. Otherwise, point $i$ is the exemplar for the cluster to which $k$ is assigned (Frey et al., 2007). Self-similarities $\mathbf{S}_{k,k}$ can be set to different values to preference some datapoints over others as exemplars, or all self-similarities can be set to the same number, in which case a larger number of results in more clusters (Frey et al., 2007). An advantage of AP clustering is that it does not require that the desired number of clusters is known in advance (Frey et al., 2007).

To address data sparsity, we test the AP algorithm not only on similarity matrices constructing from distances between data points, but also on a similarity matrix based on shared neighbors. The Jaccard similarity index is calculated by finding a set of *k* nearest neighbors for each cell and calculating the shared proportion of nearest neighbors. Because of this, the Jaccard coefficient may be a better similarity metric than the negative Euclidian and Manhattan distances or Spearman and Pearson correlations, which we examine in the first-pass analysis. Using the Jaccard coefficient means that sparse data points can still be clustered together if they have neighbors in common. The Jaccard index outputs similarities in the range [0,1], with many values being 0. Hence, we perform several transformations of **S** and iterations over the number of neighbors *k*. We applied a linear stretch, a log transform, an inverse transform, and we also explored using the tangent and hyperbolic tangent functions. All transformations were intended to explore differentially weighting highly similar points, highly dissimilar points, or points within the range, thus allowing AP clustering to converge differently. Iterating over a different number of neighbors and trying different transformations of the Jaccard index allows the set of parameters that maximize NMI to be chosen as the optimal solution within the parameter space.

There are several qualitative methods of comparing cluster assignments for the same data. However, quantitative methods facilitate optimization and comparison. For this purpose, we chose the Normalized Mutual Information (NMI), which is used by Kozakura et al. to compare different single-cell clustering methods. The NMI scores clustering solutions based on both the accuracy of the clusters, or whether both sets of clusters place the same datapoints together, and the subdivision of clusters, or whether both sets of clusters break the data into the same number of clusters (Kozakura et. al 2017). The NMI can be any value between 0 and 1, where 0 indicates no correspondence between the two sets of clusters, and 1 indicates that the cluster sets are the same (Baren et. al 2017). We calculated the NMI for each clustering method with respect to the experimental annotations, labeled as 'true' assignments for the purpose of this experiment.

## 3 Results

### 3.1 Data filtering and quality control

Cells with fewer than 200 features and any features which were expressed in less than 3 cells were removed, and cells with high UMI counts (>10000) relative to the distribution were removed, yielding 7769 cells and 16359 genes. Next, we transform the data using Seurat SCTransform to normalize and scale features. Finally, data were transformed with PCA and UMAP reductions prior to first-pass clustering and data visualization.

### 3.2 First-pass clustering analysis

For initial data clustering, we used the Louvain method, which relies on a nearest neighbor graph and is the default clustering method employed by Seurat. The 20 nearest-neighbors of each cell were computed on the first 30 PCs, and the Louvain method converged on 19 clusters (Fig 1B), NMI=0.754, whereas the number of clusters defined by Baron et al. was 14 (Fig 1A). We also clustered the data using the K-means method, with K=14 and 10 maximum iterations. K-means yielded NMI=0.690, but successful application of this method is disadvantaged in cases where the number of clusters is unknown *a priori*. Therefore, it is not considered useful in cases where the Louvain method or AP clustering may be. Finally, we performed a first-pass exploration of AP clustering using similarity metrics not intended to address data sparsity, with different diagonalizations of the similarity matrix. Sorting all the clustering solutions by NMI yielded the best first-pass AP clustering solution (AP Pearson) to have NMI=0.642 and 49 defined clusters (Fig. 1C). Based on NMI scores and the number of defined clusters, the Louvain method outperforms both the K-means and AP clustering solutions.

### 3.3 Optimizing AP clustering with the Jaccard index

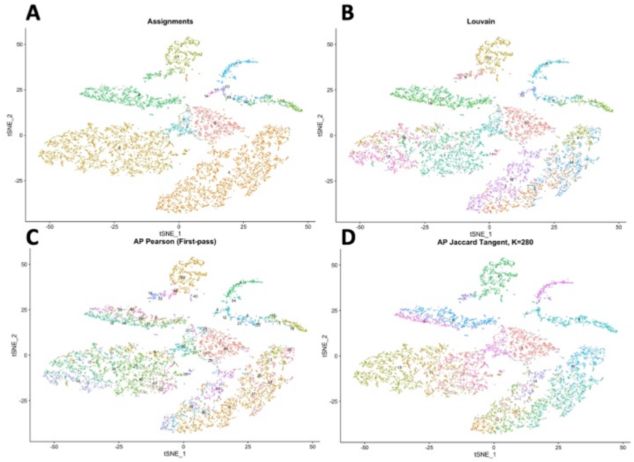
For all transformations of the Jaccard similarity index, increasing the number of neighbors *k* also increased the NMI score of the AP clustering solution. Furthermore, applying the tangent function to the Jaccard similarity matrix yielded the highest NMI scores when *k* was held constant. This could be because the tangent function pushes high similarity scores toward positive infinity and low similarity scores toward negative infinity. Hence, data points with a high Jaccard similarity score will be less likely to coalesce with lower scoring data points than when applying other functions. The hyperbolic tangent function (the sigmoid function could also have been used) flattens values both close to 0 and close to 1 while scaling values between, but this transformation was not effective. The other transformations resulted in similar NMI scores to that of hyperbolic tangent function. Finally, iterating over *k* when using the tangent function yielded a local maximum of NMI=0.695 and 14 clusters (Fig. 1D) when *k=280* (AP Jaccard). Therefore, applying tangent or a similar function with *k=280* could be the optimal parametrization for AP clustering on the Jaccard similarity index. Results from the first-pass analysis and optimization of the Jaccard index are summarized in Table 1.

**Table 1.** Benchmark results of clustering algorithms

| Method | Parameters | # Clusters | NMI |
|---|---|---|---|
| True | N/A | 14 | 1.00 |
| Louvain | 30PCs; K=20 | 19 | 0.754 |
| AP Pearson | 30 PCs | 49 | 0.642 |
| AP Jaccard | Tangent Transform, K=280 | 14 | 0.695 |
| K-means | iterations = 10; K=14 | 14 | 0.690 |

Different convergences of clustering algorithms in the gene-expression space, with respective parametrizations, show that the Jaccard similarity index was the best parametrization for AP clustering.

**Fig. 1. tSNE projection of pancreas data and cluster assignments.** There were 14 clusters in the true assignments set (A). The Louvain method (B) converged on 19 clusters



(NMI = 7.54), whereas AP Pearson (C) converged on 49 clusters (NMI = 6.42) and AP Jaccard (D) converged on 14 clusters (NMI=0.695).

Figure 1 shows that AP Jaccard appears to partition the data in the tSNE space more consistently with the assignments set than the Louvain method. This could be a reason to prefer using AP clustering over the Louvain method; however, further experiments should verify that this is consistently the case. If using the Jaccard similarity index tends to recapitulate the number of clusters better than Louvain, this may even be a better indicator of quality than the NMI. The Louvain method only outperformed AP clustering with Jaccard by 0.059, which means that further exploration of transformations on the similarity matrix could yield a superior methodology for clustering in the gene expression space. Another result of interest is that though AP Jaccard also outperformed the first-pass AP Pearson, AP Pearson had a relatively high NMI score of 0.642. Further experiments could examine these more granular clustering solutions and determine whether it is possible that subpopulations are being discovered within the data. Because of the modularity of parametrization of the AP clustering algorithm, it is also important to validate that it performs similarly on more datasets. However, this modularity is also a strength, as it provides for a more complex optimization problem than the Louvain method. More exploration into the parameter space of AP clustering is thus warranted in single cell data.

## Acknowledgements

We would like to express gratitude to the Columbia University Computational Genomics TA Philippe Chlenski and Professor Itsik Pe'er for their guidance and feedback which influenced this work.

*Conflict of Interest:* none declared.

## References

Alquicira-Hernandez, J., Sathe, A., Ji, H.P. et al. (2019). scPred: accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biol*., 20, 264.

Baron, M. et al. (2016) A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-Cell Population Structures *Cell Syst.*, 3(4), 346-60.

Braga, F. et al. (2019). A cellular census of human lungs identifies novel cell states in health and in asthma. *Nature Medicine.*, 205, 1153-1163.

Frey, J. and Dueck, D. (2007) Clustering by Passing Messages Between Data Points *Science,* 135, 972-6.

Kozakura, Y. et al. (2017) Comparison of Methods for Single-Cell Transcriptome Analysis *Information Processing Society of Japan Technical Report*, BIO-51, 1-6.

Murphy, K. (2012) Machine Learning: A Probabilistic Perspective *MIT Press.*