Anteproyecto:
Uniendo *clusters* en datos *Single-Cell*

Bernardo Álvarez del Castillo

22 de noviembre de 2019

Resumen

Los conjuntos de datos biológicos que se analizan mediante las técnicas *Single-Cell* (es decir, célula por célula) son altamente dimensionales, ya que la expresión de cada gen se contabiliza en el orden de 10^5 células (véase [1, p. 1519]). Por ello, es importante reducir al mínimo la dimensionalidad de estos datos. Una de las técnicas más antiguas, pero más empleadas para ello, es el **análisis de componentes principales** (o *PCA*, por sus siglas en inglés). Ella nos permite fijar la atención en las dos primeras componentes (cuya interpretación biológica debe analizarse luego), lo que a su vez facilita el despliegue visual de los datos, si bien indirectamente. Varios problemas se presentan inmediatamente, y en este proyecto me propongo poner en práctica algunas ideas para resolver uno de ellos en específico: la unión de *clusters* generados a través de las primeras dos componentes principales, pero que de acuerdo con cierto criterio biológico deberían conformar un mismo *cluster*. En efecto, al reducir las dimensiones de los datos originales mediante el *PCA*, nos quedamos con las componentes que presentan la mayor variabilidad (expresada a través de la varianza). Aun preservando solamente las primeras dos componentes, usualmente se mantiene un porcentaje importante de la variabilidad contenida en los datos. No obstante, alguna información adicional que no se contempla en dichas componentes podría oscurecer el hecho de que un par de *clusters* formen parte de un *cluster* más grande. El enfoque de este proyecto, por lo tanto, será analizar las siguientes componentes principales para poner a prueba esta idea.

Método

1. Dado un conjunto de datos biológico de expresión génica célula por célula, donde se manifieste la separación de *clusters* que naturalmente deberían formar uno solo, compararé su primera componente principal con algunas de las siguientes componentes por separado, analizando las gráficas de dispersión resultantes.

2. Para cada una de las reducciones de datos obtenidas en el primer paso, implementaré un algoritmo de agrupamiento, o *clusterización*, de datos previamente usado para obtener los *clusters* propuestos en algún artículo de investigación, para comprobar si los *clusters* resultantes coinciden o no con los arriba mencionados.

3. Para el mismo conjunto de datos, compararé entre sí las componentes principales que siguen a la primera, dos a dos, analizando sus correspondientes gráficas.

4. Para cada una de las reducciones de datos obtenidas en el tercer paso, implementaré el algoritmo de agrupamiento, o *clusterización*, de datos empleado en el segundo paso, para comprobar si los *clusters* resultantes coinciden o no con los del artículo, o con los que se hayan obtenido en el mismo paso.

Referencias

[1] Reyfman, P. A., Walter, J. M., Joshi, N., Anekalla, K. R., McQuattie-Pimentel, A. C., Chiu, S., ... & Verma, R. (2019). Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. American journal of respiratory and critical care medicine, 199(12), 1517-1536.