**Computational methods for trajectory inference from single-cell transcriptomics**

***Robrecht Cannoodt, Wouter Saelens and Yvan Saeys***

La identificación de la transición celular se basaba en la identificación de los marcadores en la superficie de la célula que puede distinguir un estado de otro. Esto no es tan exitoso porque por un lado es un muy de prueba y error y por otro, los estadios de transición no siempre son separables por el fenotipo.

Nuevos estados de transición se corresponden con un densos clusters de células en el continuo que se pueden detectar usando técnicas de unsupervised clustering.

El proceso de interés: La regulación del transcriptoma.

Durante su desarrollo los cambios en la actividad del factor de transcriptoma inducen: modificaciones y remodelación de la cromatina y al final un reagrupamiento diferente de la maquinaria base del trancriptoma.

Entonces, modelar la dinámica de la regulación de genes es esencial para entender mejor por qué un proceso dinámico celular progresa en varios pasos y qué es lo que se hace mal en caso de una enfermedad.

Un proceso dinámico puede ser sincrónico o asincrónico (hematopoiesis).

La idea de nuestro trabajo: single cell snapshot + TI (trajectory inference) = accurate time series dataset para cada población de células. Usando la data single cell **nos ahorramos de tener que estudiar los factores de transcripción. [17,30,76,77][Computational methods for trajectory inference from single-cell transcriptomics, Robrecht Cannoodt, Wouter Saelens and Yvan Saeys].**

El poder de single cell: poder identificar transiciones exóticas evadiendo overfitting on outlier cells. Va a servir para resolver tanto trajectorias controversiales como ya establecidas en el sistema inmune y por debajo.

**Manifold learning-based mathods for analyzing single-cell RNA-sequencing data**

***Kevin R Moon et al***

Manifold: Es una construcción matemática que representa a una espacio liso variable en el sentido local euclideano. Es parecido a una hoja 2D lisa en un espacio 3D. La hipótesis de manifold se usa para modelar la data como una colección vecinos locales que varían lisamente (sin hacer agujeros) que son de baja dimensionalidad con respecto que el espacio ambiente. Éstas propiedades suelen ser ciertas para grados altos de data single-cell dado que las células tienden a esparcirse de forma lisa (smooth) y que el espacio es intrínsecamente de baja dimensionalidad con lo que respecta a un tema en particular o proceso.

También dice que en el espacio celular típicamente las transiciones son smooth donde las células van haciendo transiciones en pequeños pasos en vez de forma dramática.

Es importante que en la hipótesis de manifold el ruido en los datos es tratado como una perturbación de gran dimensionalidad por fuera del manifold y que se puede eliminar el ruido proyectando los puntos en el smooth manifold.

Pasos:

* El denoising tiene que ver con los dropout. Esto lo explica en pag 37. También lo une con las técnicas que vi en los artículos anteriores de reducción de dimensionalidad.

* Estudiar la pairwise gene interaction usando por ejemplo mutual information.
* El pseudotime y el graph fitting. La inferencia de la trayectoria del pseudo tiempo decribe el eje latente de desarrollo del estado de transición de la célula. Ésta organización del espacio de la célula hace que se enfaticen las variaciones smooth de las poblaciones experimentales.

En criollo la idea es que en la estructura celular hay una manifold y entonces, al aplicar éstas técnicas encontramos la estructura topológica de manifold subyacente y a partir de eso por la idea de distancias euclideanas podemos hacer la inferencia de la trayectoria del pseudo tiempo.

Es interesante porque afirma que varios usan como reducción de la dimensionaldiad el PCA pero ésta técnica tiene dos hipótesis que en general los procesos de regulación celular no cumplen…(pag 41) Y el framework de manifold learning es útil para la reducción de dimensionalidad como la visualización.

Lo más importante es que sugieren que manifold learning modela la data single-cell de tal forma que permite visualizar varios sistemas biológicos. Da el ejemplo: usa para la progresión de un single branch en el desarrollo de células T (las que están involucradas en la mediación celular inmunológica caracterizados por el receptor T en la superficie de la célula) y en la diferenciación de progenitores myeloid y erythroid en hematopoietic steam cells (hematopoiesis es la producción de los componentes celulares de la sangre. Se generan nuevas células de la sangre en una persona todos los días) [cita 63 del paper].

Hay que tener en claro que son varios los pasos que hay que hacer. En general, se tiene la matriz de genes y células donde veo qué gen está prendido en cada célula. Luego, identifico la estructura de los vecinos de una forma de aprendizaje y recuperar la representación latente de baja dimensión.

Cómo hace esto?: Bueno, luego de la reducción de dimensionalidad a algo 2D la distancia entre los puntos es calculada y usa una función kernel que calcula la afinidad por las distancias. Sólo se preservan las relaciones locales y por último, se calcula el manifold que se encuentra por debajo.

Cómo entra lo del pseudotiempo?: Cuando ya tenemos modelado el mecanismo de la generación de la data de un dataset entonces, el eje en el manifold de mayor cambio se lo calcula como pseudotiempo que va a representar la progresión o diferenciación en desarrollo. Los estados de alta densidad se usan para clustering y el ruido es modelado como una perturbación al manifold y removido al proyectar nuevamente la data al manifold. Así, al final del día, la representación del manifold se usa para la reducción de dimensionalidad y la visualización.

A ver sí entiendo bien. La idea es aplicar la idea de manifold una vez para llegar algo donde puedo modelar el proceso biológico (o sea ver una estructura subyacente) para después encontrar cuál es el eje de mayor cambio y decir bueno, ese es el pseudotiempo. Entonces volvemos a la expresión de cada gen (nos salimos de la reducción de dimensionalidad) y tenemos una expresión de cada gen en función de éste pseudotiempo y a partir eso encontramos los estados de mayor densidad para hacer clustering. Luego le quitamos el ruido en gráficos de un gen versus otro y devuelta hacemos esto de manifold para la reducción de dimensionalidad + visualización. (Fig 1)

Una forma rápida de pensar al manifold es decir que le saca el ruido a todo y se queda sólo con las relaciones de bajo alcance entonces,podemos ver el proceso biológico subyacente. Es un algortimo que hace eso y acompañado de demás cosas, como digo en el párrafo anterior, se pueden obtener buenas conclusiones. Sin embargo, manifold no entra en la idea de la dinámica de la red. Para eso por ejemplo hay que usar mutual information (MI).

**Learning regulatory models for cell development from single cell transcriptomic data**

***Ann C. Babtie et al***

Complex gene regulatory network (GRN): Comprende la activación y la represión de interacciones entre factores de transcripción y sus targets de control de los estados de transcripción de las células. En un framework dinámico éstas redes son vistas como que regulan la probabilidad de que una célula ocupe diferentes estados de la expresión genética. En crillo, una célula tiene la posibilidad de tener muchos genes pero, no todos están prendidos. Sino que va a depender de la función de esa célula cuáles van a ser los genes prendidos. Además esto evoluciona en el tiempo, algunos genes que no estaban prendidos pueden prenderse en un futuro y viceversa (está bien?). Estados atractivos estables están asociados con tipos discretos de células que se observan experimentalmente y el paisaje potencial determina la ruta de transición probable entre estados [9-11]. Al tener experimentos single-cell en vez de tener transiciones smooth son discontinuos y estocásticos los eventos de transiciones. Se ve la heterogeneidad.

Pregunta: El orden de los pasos que muestra en Fig.1 pareciera ser distinto al orden del manifold del artículo anterior. En la pag.76 plantea usar el PCA o el MI para la inferencia de la estructura de la red mientras que en el artículo anterior pensé que era para la reducción de dimensionalidad.

Algo interesante que propone en las conclusiones es diseñar experimentos para testear y refinar nuestras hipótesis y verificar las conclusiones que obtenemos en in vitro data que se corresponden con observaciones in vivo.

Algo que está bueno para el futuro es que habla de usar tipos de datos heterogéneos para obtener conclusiones más profundas sobre la biología involucrada.