SLICER

Joshua D. Welch, Alexander J.Hartemink, Jan F. Prins

Septiembre 2018

1. Introducción

El objetivo es inferir un manifold de baja dimensionalidad que está embebido en un espacio de alta dimensionalidad. La idea es que éste manifold capture las relaciones geométricas entre las células [1,2].

Hipótesis de éste modelo: La diferencia significativa entre las células es dónde se encuentran durante el proceso biológico. Entonces, la secuencia de la expresión de genes se encuentran en una trayectoria en el espacio de alta dimensionalidad.

Técnicas previas: Monocle (ICA) y Wanderlust.

¿Qué es SLICER?: Selective locally linear inference of cellular expression relationships. Emplea la técnica de locally linear embedding (LLE) para reconstructir trayectorias celulares.

¿Cuáles son las ventajas que presenta?:

- 1. Hace una selección de genes para construir la trayectoria celular sin necesidad de conocimiento biológico previo.
- 2. Es una técnica no lineal de reducción de dimensionalidad que permite observar las relaciones no lineales entre los niveles de las expresiones de los genes y la progresión a lo largo de proceso.
- 3. Emplea una métrica, geodesic entropy, que detecta automáticamente el número y la ubicación de los branches en la treyectoria celular.
- 4. La capacidad de detectar unos features nuevos, "bubbles".

2. Pasos del software

PRIMERO. SELECCIÓN DE GENES. Primera parte de la fig. 2 a.

Toma como input la matriz de los niveles de expresión de los genes. Calcula una magnitud "neighborhood variance"para seleccionar aquellos genes que van a ser usado para construir la trayectoria. Es un método para remover los genes que muestren una fluctuación random y elige sólo a aquellos que varían incrementalmente de una célula a otra de forma sistemática y gradual.

Los puntos que estan juntos en un espacio euclideano tienen una tendencia a que permanecer juntos en el manifold. Entonces, podemos usar la simulitud de los genes en los puntos vecinos para aproximar los cmabios en el gen moviendose a lo largo de la trayectoria.

Entonces, se calcula para un dado gen g por un lado la σ_g^2 varianza con respecto a todas las muestras y por otro, la "neighborhood variance"que intuitivamente es una varianza computada en vez de con

respecto a la media con los puntos vecinos y cuantifica cuanto g varia con respecto a los vecinos en cada muestra:

$$S_g^{2(N)} = \frac{1}{nk_c - 1} \sum_{i=1}^{n} \sum_{j=1}^{k_c} (e_{ig} - e_{N(i,j)g})^2 \tag{1}$$

donde e_{ij} es el nivel de expresión del j-ésimo gen en la muestra i-ésima. N(i,j) son la cantidad de vecinos del gen j-ésimo gen en la muestra i-ésima y k_c es el mínimo número de vecinos para tener un gráfico conectado.

El criterio de selección es:

$$\sigma_g^2 > S_g^{2(N)} \tag{2}$$

SEGUNDO. SELECCIÓN DE VECINOS. Segunda parte de la fig. 2 a.

Se toma un número k de vecinos cercanos que son aquellos que forman una forma lo más cercana a posible a la de una trayectoria para construir un embedding de baja dimensionalidad. Esto lo mide por el alpha convex hull del embedding.

Para un dado k realiza LLE y calcula l, el camino corto mas largo y toma $\alpha = l/10$, me quedo con un 10 porciento de los puntos y calculamos el area a del "$\alpha hull$"($\alpha eselradio$).$Sedefineelwidthdelembeddingcomo$w a/l que cuantifica cuanto el embedding se parece una trayectoria. Se toma el k tal que:

$$k = argmin_k(w_k) \tag{3}$$

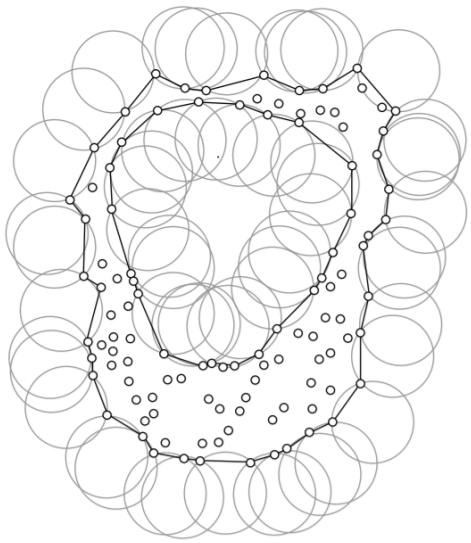

Figura 1

TERCERO. Primera parte de la fig. 2 b.

A partir del paso segundo tenemos un k-nn graph de alta dimensionalidad y por medio del locally linear embedding se proyecta al set que obtuvimos en el paso segundo en un espacio de menor dimensionalidad.

Se tiene el matriz $E_{nxm} = (e_{ij})$ de la expresion del gen j en la muestra i. Y la del low dimensional embedding L_{nxd}. Ellos dicen que en general $d = 2$ es una decision razonable.

Se realizan dos pasos. Primero un set de pesos de reconstruccion es aprendido, W_{nxd}, para que cada punto en el espacio de alta dimensionalidad sea representado como una combinacion lineal de sus

k-nearest neighbors.

$$W = argmin_w \sum_{i=1}^{n} |E_i - \sum_{j=1}^{k} w_{ij} E_j|_2^2 \tag{4}$$

Y luego, con este valor del W lo que hacemos es calcular nuestra matriz de baja dimensionalidad.

$$L = argmin_L \sum_{i=1}^{n} |L_i - \sum_{j=1}^{k} w_{ij} L_j|_2^2 \tag{5}$$

CUARTO. Segunda parte de la fig. 2 b.

Ahora usamos ese low-dimensional embedding para construir otro gráfico de vecinos donde las células son ordenadas basadas en el camino más corto desde una célula específica.

Como segundo paso lo que se hace es aplicar el logaritmo de Dijkstra para hallar el camino mas corto dado un starting point en el sentido euclideano. Eso se aplica sobre el grafo de L.

QUINTO. Fig. 2 c.

Se calcula una métrica que se llama entropía geodésica para poder detectar la presencia, el número y la ubicación de los branches en la trayectoria celular.

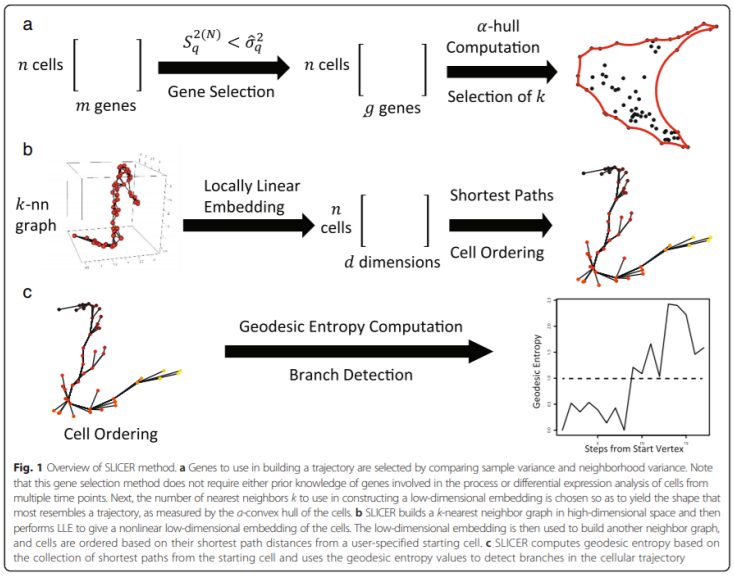

Fig. 1 Overview of SLICER method. **a** Genes to use in building a trajectory are selected by comparing sample variance and neighborhood variance. Note that this gene selection method does not require either prior knowledge of genes involved in the process or differential expression analysis of cells from multiple time points. Next, the number of nearest neighbors *k* to use in constructing a low-dimensional embedding is chosen so as to yield the shape that most resembles a trajectory, as measured by the *α*-convex hull of the cells. **b** SLICER builds a *k*-nearest neighbor graph in high-dimensional space and then performs LLE to give a nonlinear low-dimensional embedding of the cells. The low-dimensional embedding is then used to build another neighbor graph, and cells are ordered based on their shortest path distances from a user-specified starting cell. **c** SLICER computes geodesic entropy based on the collection of shortest paths from the starting cell and uses the geodesic entropy values to detect branches in the cellular trajectory

Figura 2

3. Comparación con métodos previos

Se emplean grupos de funciones que simulan 5 "pathways"distintos. se toma que los genes simulados por una familia son análogos a los genes co regulados en un "pathway"biológico donde todos cambian como una respuesta a un mecanismo regulatorio común. Le mete ruido haciendo un reshuffle donde de forma aleatoria hace una permutación entre los valores de algunos genes regulado por el parámetro p.

3.1. Accuracy

Proponen un medida: "percent sortednessçomo una métrica para definir que tan bueno es el método. Se calcula la proporción relativa de cuántos puntos son compartidos entre el orden computado por

SLICER y el que se conoce, de antemano, que describe el sistema.

3.2. Elección de los k vecinos para la construcción de la trayectoria

En el paso 2 el software lo que hace es elegir los k para armar el k-nn graph en el espacio de alta dimensionalidad empleando α convex hull approach. Para poder afirmar apropiadamente de que es una "buena"decisión de los k lo que hacen es tomar distintos valores de k en $[5, 10, ..., 45, 50]$ y quedarse con el valor de k para el cual tiene una mejor accuracy (un percent sortedness mas alto). Y entonces, compara SLICER best k con SLICER auto k.

3.3. Capacidad de detectar branches y "bubbles"

Lo que hacen es simular una trayectoria que en un solo camino se divide en dos branches que luego vuelven a converger en un único camino. Esto es lo que se llama "bubble".

Entonces, comparan el manifold que observa LLE embedding y SLICER. El primero lo que hace es reconocer el bifurcación de los caminos y como estos se vuelven a unir mientras que el segundo representa otro patrón que es característico de un bubble.

Por otro lado, también testean que tan bien detectar branches SLICER frente al ruido. Define la siguiente métrica: El porcentaje relativo de las células asignadas correctamente al branch.

4. El potencial de la herramienta

Los autores emplean el software para estudiar el desarrollo de celulas de cancer de pulmon en ratones. Se tienen muestras single cell en los dias embrionicos 14.5, 16.5 y 18.5 y el dia posnatal 107. Entonces, ven el desarrollo de celulas progenitoras, intermedias y celulas comprometidas en specialized cell fates.

El método selecciona aquellos genes cuya varianza del nivel de expresion excede la "neighborhood variance". Asi produce una lista de 660 genes. Luego, hace el 2 dimensional embedding de la data usando LLE. Despues se elige una celula como la **celula inicial** y construye el nearest neighbor graph en baja dimensionalidad. Y encuenta el "single-source shortest pathsüsando el algoritmo de Dijkstra.

Se obtiene un resultado que es genérico, o sea que no diferencia en el tipo de celula, que se en Fig. 3. Luego, para identificar cuales son las celulas que estan involucradas en cada uno de los procesos. Se hace uso de los marcadores que se emplean en el experimento de laboratorio donde a partir de ellos se obtienen los graficos que se ven en Fig. 4.

Entonces, también hace un uso de la entropía geodésica y encuentra el valor de la cantidad de pasos en el cual el valor de la entropía supera al 1 por lo que se hacen los branches ya identificados. Entonces, afirman que esa ubicación es un punto de decisión en la diferenciación celular ya que la célula o bien procede hacia el AT1 cell fate o el AT2.

Entonces, en definitiva lo que se termina teniendo es un mapa donde se ve la evolución en vez de en el tiempo en la biología, de células tempranas embrionarias en dos tipos, alveoales tipo 1 o de tipo 2 en Fig. 5.

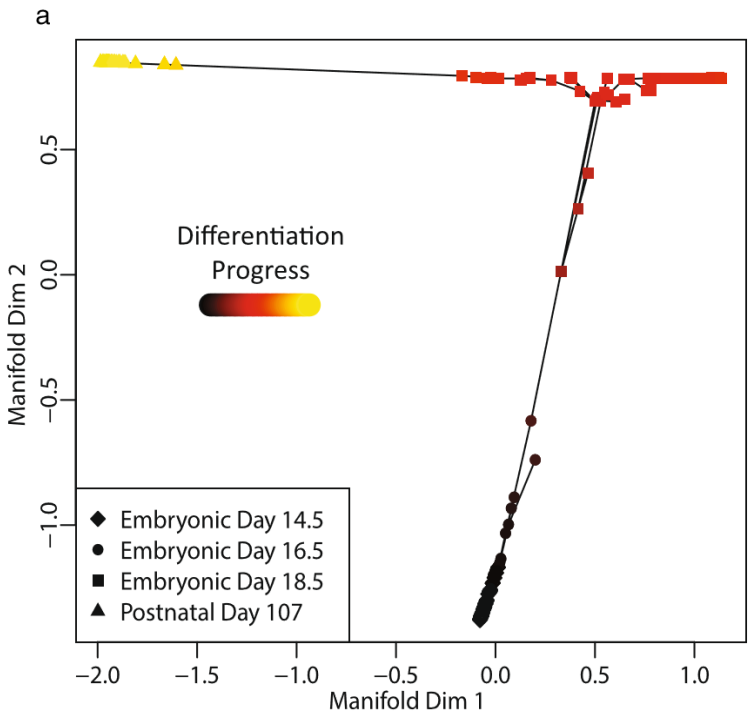

Figura 3

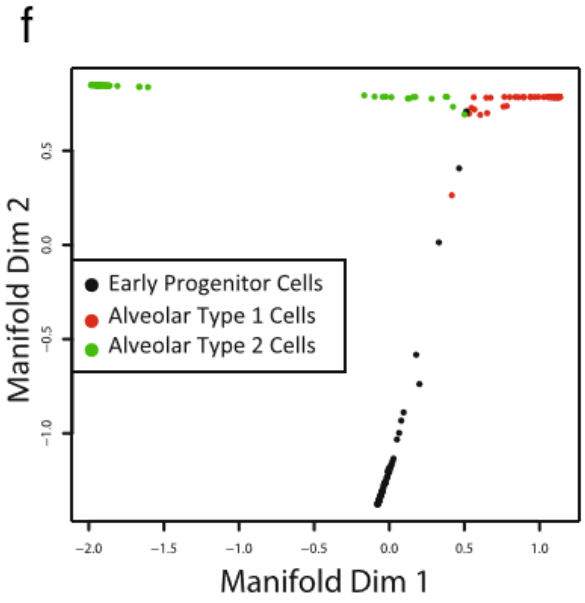

Figura 5

5. Trabajos realizados

Querían ver los cambios en la expresión genérica durante la neural stem cell activation después de una lastimadura en el cerebro. Tenían 271 células con más de 1000 genes.

PASO 1: 1000—-661 genes (variance y neighborhood variance).

PASO 2: Detectar clusters en el low-dimensional k-nearest neighbor graph. Estos son asociados a tipos de células. Por ejemplo, acá se dieron cuenta que un tipo de célula era de oligodendrocytes que se dio por un overlap con los marcadores para aislar las neural stem cells. Entonces, uno debería quitar todas esas células para ver la trayectoria.

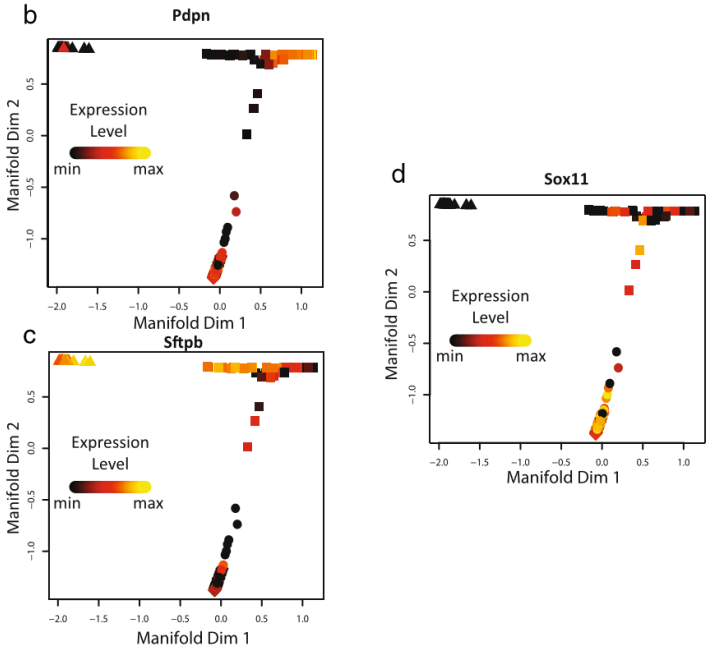

Figura 4

PASO 3: Estudiar la expresión de marcadores conocidos de genes. Entonces, cada uno de los clusters van a terminar asociados a estos tipos de celulas.

PASO 4: Estudiar la entropía geodésica para encontrar los branches con precisión. Poder decir: Hay un branch a los 50 pasos (desde la starting cell) que tiene que ver con la distinción entre NSCs y neuroblast. Luego, se da otro que es de la separación entre neuroblast y oligodendrocytes.

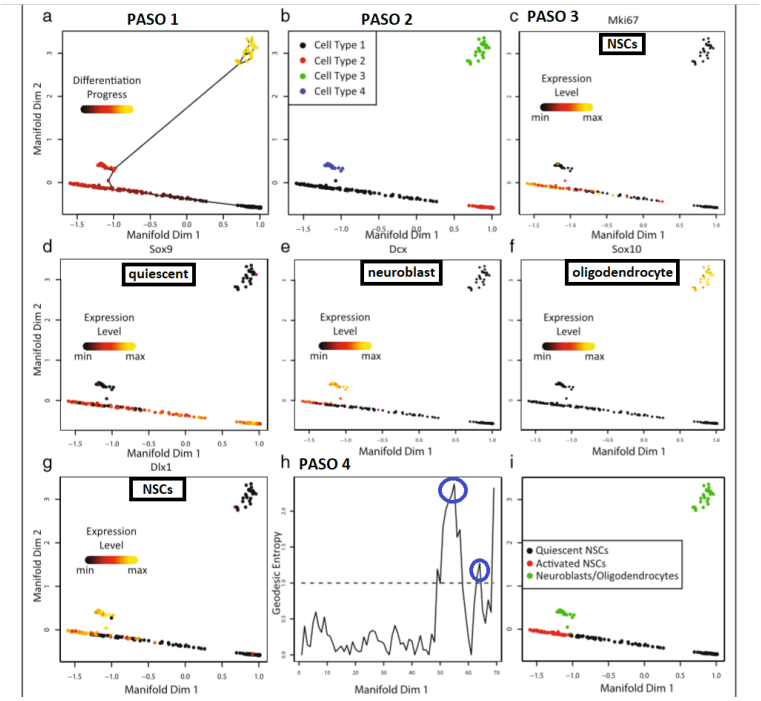

Figura 6

Entonces, como resultado del trabajo lo que se obtiene es poder relacionar la evolucion en el tiempo con los cambios en la expresion genetica. Entonces, aca esta apagado el NSC y luego se divide en dos, en aquellos prendidos o neuroblast/oligodendocytre (el branch) y luego se da una diferenciacion entre estos dos ultimos. Asi evolucionan en el tiempo la activacion de neural stem cell despues un accidente cerebral.

6. Buena comparacion

Usan 5 funciones f(t) que las multiplican por variables distribuidas normalmente c_i.$Asitienenlasmatricesdec$ 0.1, ... , 80 como valores de t (801). Para generar ruido lo que hace es permutar algunos puntos entre unas trayectorias a otras (de las 5 funciones). Toma un set de 5 genes (uno de cada funcion) y hace el reshuffle con una probabilidad fija, p.

6.1. Cuantificar la performance

Comparamos el orden que obtenemos con SLICER con el orden que conocemos de las funciones. Esto lo cuantifican con la percent sortedness.

$$(1 - \frac{s}{\binom{n}{2}}) * 100 \tag{6}$$

donde s es el numero de pares que se quedan afuera del orden.

7. Posibles cosas a hacer

- Factores como el número de branches, el tamaño relativo de éstos y el alcance de los cambios por sobre el set de células depende de la cantidad de células analizadas con ésta técnica single-cell. Podría ir variándolo e ir viendo cómo eso afecta.

8. PREGUNTAS

- Cómo es que obtengo los set de datos donde por un lado uno es algún determinado marcador y por otro son todas las células.
- El caso que muestran en trabajos realizados veo que hablan de tipos de celulas. No entiendo donde es que entra la idea de cuales son los genes prendidos o apagados.
- El significado fisico de las ec. 4 y 5.
- como es que hago un grafo a partir de esa matriz L. Tomo cada fila (seria para una dada muestra a dado tiempo?) y hago el grafo y arriba de ese el resto de las filas?
- como afecta que sean co-expressed genes para la seleccion.