Reconocimiento de Patrones Genéticos por Medio de Grafos

Cristóbal R. Santa María*, Romina Rebrij**Victoria Santa María***, Luis López*y Marcelo Soria****
*DIIT-UNLaM, **Hospital Italiano (CABA), ***Instituto Lanari-FMed-UBA, ****FAUBA Florencio Varela 1903 San Justo Pcia. de Buenos Aires 54-011-44808952
csanta_maria@ing.unlam.edu.ar
rominarebrij@gmail.com
vctrsntmr@gmail.com
llopez@ing.unlam.edu.ar
soria@agro.uba.ar

Resumen

Se expone la línea de investigación que se lleva adelante en el Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLaM. Se detallan resultados del proyecto de investigación "Aplicación de Técnicas de Data Mining para Análisis del Microbioma Humano según Funcionalidades Metabólicas", C200 del Programa de Incentivos. Con él se intenta aportar procedimientos para analizar la relación clínica entre el microbioma intestinal y la presencia de patologías. Esto comprende la obtención de muestras de microbiomas de pacientes, la identificación funcional de las secuencias genéticas y la determinación de la distribución de frecuencias por especies en cada paciente. En el proyecto de investigación anterior C169 se habían obtenido datos de secuencias del gen marcador 16S rRNA. La necesidad de establecer ahora una clasificación por funcionalidades metabólicas para todos los genes presentes en cada microbioma, llevó a la búsqueda de nuevos datos crudos (no ensamblados) y al análisis de los procedimientos de extracción, control de calidad, limpieza y ensamble.

Palabras Clave: Microbioma Secuencias Ensamble Grafos

Contexto

El cuerpo humano es colonizado por una comunidad de microorganismos que se denomina microbioma y contiene diez veces más células que las suyas propias. La cantidad de genes presentes en total es varios órdenes de magnitud mayor que la del genoma humano. La nueva generación de tecnologías de secuenciación de ADN ha permitido estudiar las características del microbioma humano. El objetivo de estos estudios metagenómicos es analizar la estructura y la dinámica de las comunidades, para establecer cómo se relacionan sus miembros entre sí, cuáles son las sustancias que producen y consumen, y cómo se modifica la comunidad en presencia de enfermedades. El estudio por medio de la asignación funcional de cada gen del microbioma y su ubicación dentro del complejo de actividades metabólicas que ocurren en el paciente hospedador, en la comunidad microbiana, y en la interacción entre ambos, busca reconocer actividades metabólicas en el paciente asociadas con la presencia de enfermedades. Este es un campo de investigación muy activo con proyectos como el Metagenomics of the Human Intestinal Tract (MetaHIT), y abarca desde los aspectos médicos hasta el desarrollo y aplicación de nuevos algoritmos de explotación de datos y

reconocimiento de patrones. El objetivo general es entender el funcionamiento del microbioma humano a partir del procesamiento y análisis de muestras de secuencias de ADN, y construir nuevas herramientas de software para caracterizar el curso de patologías.

Introducción

El trabajo consiste en obtener los datos secuenciados de una muestra compuesta por varios microbiomas. El conjunto de secuencias ya ensambladas del microbioma habrá de compararse con otra base de datos de funciones genéticas para agrupar los genes integrantes por función y así obtener la distribución de frecuencias según las funciones metabólicas que las secuencias integrantes revelan [1]. Una vez formadas las matrices que representan por fila las distribuciones de cada microbioma individual estos pueden agruparse en clusters. Cada fila del conjunto representa a un paciente y en la base de datos esa instancia es un vector donde cada componente corresponde al número de genes del microbioma identificados con una dada función metabólica. Las características de cada agrupamiento logrado deben cotejarse con las apreciaciones clínicas de los pacientes que lo integran, ya obtenidas por otras vías diagnósticas, para apreciar el punto hasta el cual resultan útiles en la evaluación médica de la patología investigada.

Líneas de Investigación, Desarrollo e Innovación

En esta línea de trabajo ya se han estudiado procedimientos de obtención de datos y clustering de lo que se ha dado cuenta en presentaciones anteriores. En esta oportunidad se analizó el trabajo completo que se efectúa sobre las secuencias desde que son obtenidas por el secuenciador hasta que se construyen los "contigs" mediante el ensamblado de secuencias que se realizó aplicando grafos. Los datos utilizados son nuevos pues se buscó que sirvieran luego para la identificación funcional. Su origen es la tecnología de secuenciación Illumina, y la forma de ensamblado es "de novo", es decir sin que se utilicen genomas preexistentes y anotados como guía para ensamblar. Esta parte del trabajo se incluye dentro de los objetivos más generales que se intentan alcanzar en esta línea de investigación, desarrollo e innovación:

-Dominar la tecnología de almacenamiento, comparación y distribución funcional según las secuencias obtenidas del microbioma intestinal

-Determinar los métodos computacionales convenientes para los agrupamientos de microbiomas que revelen sus características clínicas.

-Establecer algoritmos de predicción entrenados y testeados para la evaluación clínica.

-Establecer una "pipeline" para la aplicación a pacientes locales

-Obtener muestras propias, enviarlas a secuenciar y aplicar los procedimientos probados.

Resultados y Objetivos

Los datos de secuenciación de este estudio son públicos y se busca poner a punto y automatizar, las operaciones bioinformáticas necesarias: selección de software más adecuado para cada una de dichas operaciones, análisis de calidad de las secuencias, diseño de una metodología de limpieza de las secuencias, validación, ensamblado de los metagenomas e integrar todos estos pasos.

El estudio cuenta con 143 muestras que se distribuyen de la siguiente manera: 16 de endoscopias, 99 de materia fecal y 28 de hisopados rectales. Las secuenciación de

ADN total se realizó con tecnología Illumina con una estrategia de "paired-ends" de 300 nucleótidos, por lo que cada muestra está compuesta por dos archivos de secuencias. Esto significa que para cada fragmento de ADN analizado, se secuencian 300 nucleótidos desde cada extremo.

Todos los pasos se realizaron corriendo el sistema operativo Linux, distribución Ubuntu 16.04. Se procedió a la descarga de los 286 archivos utilizando la herramienta SRATOOLKIT de NCBI y se eliminaron dos muestras (4 archivos) que contaban con muy pocas secuencias. Las muestras restantes tenían entre, aproximadamente, 41600 y 521000 bases.

Se realizó el control de calidad con el software FastQC y se determinó que casi todas las secuencias tenían restos de dos de los adaptadores que usa Illumina para la secuenciación, uno en las secuencia F ("forward") y otro en las secuencia R ("reverse") de cada "paired end". Además se determinó que las frecuencias de cada base en los primeros 15 nucleótidos de las secuencias presentaban un nivel de variabilidad muy alto, que no era compatible con lo que se observaba más adelante y debido, posiblemente, a algún artefacto de la secuenciación que generaba "ruido". También la calidad promedio de las secuencias caía por debajo del valor umbral que se fijó en 25 a partir de una posición que variaba para cada secuencia, pero que en general se ubicaba después de la posición 240. Para determinar el tipo exacto de secuencia contaminante y para tener una información más precisa del lugar donde ocurría se utilizó el software SCYTHE [3]. El proceso de limpieza se realizó con el programa CUTADAPT que efectúa la limpieza de adaptadores, cortes por caída en los valores de calidad, eliminación por largo mínimos, cortes en posiciones arbitrarias, etc. En primer lugar se realizaron pruebas preliminares para determinar las opciones específicas de limpieza y los valores óptimos de los parámetros del programa. El proceso definitivo se efectuó en dos pasos. En el primero se eliminaron las secuencias contaminantes, se eliminó la parte 3'de las secuencias que presentaran una caída en su calidad por debajo del valor umbral 25 mencionado antes y si alguna de las secuencias de un par "paired-end" después de estos cortes resultaba con una longitud menor a 50 bases se procedía a eliminar el par completo. En el segundo paso se eliminaron los primeros 15 nucleótidos del extremo 5' y se volvieron a filtrar los pares para eliminar a aquellos con al menos un miembro de longitud menor a 50 bases.

Después de la limpieza se volvió a revisar la calidad de las secuencias con FastQC, con resultados satisfactorios.

Con las secuencias limpias y filtradas se procedió al paso de ensamblado. En la secuenciación el genoma es fragmentado. La longitud de cada fragmento depende de la tecnología empleada (≈600 bp con Illumina en paired end) y se necesita ensamblar los fragmentos luego de la secuenciación para identificar los genes de un individuo. La longitud en pares de bases (nucleótidos) de cada genoma varía según el tipo de organismo. Se entiende por read a cada fragmento obtenido del secuenciador. Varios reads se ensamblan para formar un contig que expresa los alineamientos de los nucleótidos que se encuentran en ellas.

La intención es que los contigs reconstruyan una parte de cada gen presente en la cadena de ADN que fue fragmentada [5] según se ve en la figura

```
-- -ACCGT--
-- ----CGTGC
 --TTAC-----
- -TACCGT--
----------
--TACCGTC
```

En 1735 L. Euler resolvió el llamado problema de los siete puentes [5]. Partiendo de una cualquiera de cuatro regiones, conectadas por siete puentes, en que quedaba dividida la ciudad, determinó las condiciones en las que cualquier problema similar puede resolverse afirmativamente. Euler demostró que:

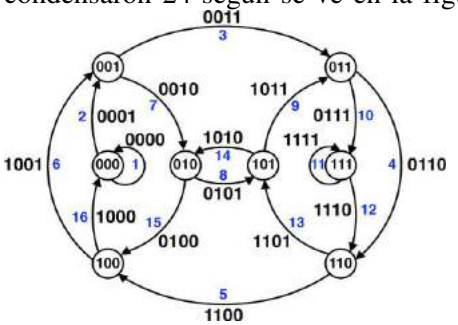
i) Si hay más de dos regiones a las cuales lleva un número impar de puentes entonces el camino buscado no existe

ii) Si hay solo dos regiones unidas por un número impar de puentes el camino puede hacerse iniciándolo en cualquiera de ellas.

iii) Si no hay ninguna región a la que lleva un número impar de puentes el camino siempre podrá hacerse iniciándolo en cualquier región. Modernamente se diría que para tratar el problema Euler representó cada puente con un arco del grafo y cada región con un nodo. Un ciclo euleriano comienza y termina en un mismo nodo pasando una sola vez por cada arco

En 1946 Nicolaas de Bruijn buscó resolver el problema llamado de la "supercadena": encontrar la supercadena de caracteres más corta que contuviera a todas las posibles subcadenas de k símbolos (k-mers) de un alfabeto de n símbolos.

En un alfabeto de n caracteres existen subcadenas de k símbolos. Si los símbolos del alfabeto fueran las letras de los nucléotidos del ADN habría 4^3=64 subcadenas posibles de tres nucleótidos (3-mers). Si en cambio el alfabeto estuviera formado por los símbolos binarios 0 y 1 todos los posibles 3-mers serían 000 001 010 011 100 101 110 y 111. Así se ve que la supercadena 0001110100 contiene a todos estos 3-mers. Es decir con 9 símbolos se condensaron 24 según se ve en la figura.



De Bruijn imaginó esta solución por medio de grafos teniendo en cuenta los resultados de Euler [7] Si, por ejemplo, el alfabeto está formado por los símbolos 0 y 1 (n=2) y se quiere hallar la supercadena circular más corta que contenga a todos los k-mers con k=4, basta considerar los (k-1)-mers (k-1=3) como los nodos de un grafo cuyos arcos dirigidos se constituyen tomando el prefijo del nodo de partida y el sufijo del nodo de llegada. Los distintos arcos así formados constituyen un ciclo euleriano. La supercadena cíclica resultante está formada por cada uno de los prefijos de los arcos. Es decir: 0000110010111101 A cada nodo salen y llegan en suma un número par de arcos con lo que se está en la condición iii) del teorema de Euler y por lo tanto el ciclo euleriano existe y es único.

En principio aplicar grafos al ensamblado de secuencias implicaría representar cada read por un nodo y los solapamientos entre lecturas por arcos.

Las técnicas de ensamblado de ADN utilizan un valor de k≈55 para fragmentos obtenidos por tecnología Illumina. Se pueden organizar las superposiciones de estas subsecuencias haciendo coincidir el subfijo de una inicial con el prefijo de la otra final. Esto se expresa en un grafo cuyos nodos son los k-mers y los arcos sus superposiciones. El ciclo hallado es Hamiltoniano pues pasa solo una vez por cada nodo.

10^6=1000000 es un número de lecturas que podría generar la secuenciación Illumina. Ellas requerirían unos 10^{12} alineamientos de k-mers y si hubiera 10^9 lecturas serían necesarios 10^{18} alineamientos. No existe algoritmo eficiente para esto pues la cantidad de operaciones a realizar no podría efectuarse en tiempos polinómicos. Como hallar un ciclo hamiltoniano es un problema NP-completo el proceso de ensamblado de ADN enfrenta a la computación con sus límites teóricos actuales.

En este punto entran Euler y de Bruijn Resulta más sencillo y resoluble computacionalmente encontrar un ciclo que pase una sola vez por cada arco. Euler probó que si el grafo es conexo y no dirigido contiene exactamente un ciclo (euleriano) cuando cada nodo del grafo se asocia a un número par de arcos que lo conectan con otros nodos. En un grafo dirigido la cuestión es análoga: el número de arcos que parten del nodo tiene que ser igual al número de arcos que llega a él. (Grafo balanceado). En particular los grafos de de Bruijn contienen entonces un ciclo euleriano. Se trata de hallarlo.

No todos los supuestos de de Bruijn se cumplen en el caso de la fragmentación del genoma. En primer término no necesariamente se presentan todos los k-mers que podrían formarse. La solución que se ensaya es tomar k≈55 con la esperanza de que todos los arcos posibles estén efectivamente en el grafo. En segundo lugar algunos k-mers pueden repetirse frecuentemente. Para resolverlo se establece la "multiplicidad" del k-mer. Si su multiplicidad es m, se conecta su prefijo con su sufijo m veces y como el grafo también resulta balanceado existe el ciclo euleriano. Además, en general, la tecnología Illumina puede producir errores de lectura que, en tal caso, se intentan corregir antes del ensamblado. Existe software libre que puede usarse para realizar el ensamblado de secuencias por esta vía. Por ejemplo el IDBA.UD desarrollado por el el departamento de ciencias de la computación de la Universidad de Hong Kong [8] y [9]

Formación de Recursos Humanos

En el equipo de trabajo participan un magister en explotación de datos y otra en bioinformática, un doctor en biología, una médica, 2 ingenieros en sistemas, una matemática y un estudiante de ingeniería informática. Está en curso una tesis de doctorado y otra de maestría.

Referencias

[1] Arumugam, M et al. Enterotypes of the human gut microbiome. Nature 2011 may 12; 473(7346): 174-180. doi:10.1038/nature09944

[2]https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

[3] https://github.com/vsbuffalo/scythe

[4](https://cutadapt.readthedocs.io/en/stable/

[5] Lander ES, Waterman MS (1988). "Genomic mapping by fingerprinting random clones: a mathematical analysis". Genomics. 2 (3): 231–239. doi:10.1016/0888-7543(88)90007-9. PMID 3294162.

[6] Euler, L. Commentarii Academiae Scientiarum Petropolitanae 8, 128-140 (1741)

[7] Campea, Ph E C; Pevzner, P A; Tesler, G. How to apply de Bruijn graphs to genome assembly. Nature Biotechnology Volume 29 Number 11. 987-991. (2011)

[8] Peng, Y; Leung, H C M; Yiu, S M; Chin F Y L. IDBA-UD: de novo assembler for single-cell and metagenomic sequencing data with hibhly uneven depth. Bioinformatics. Vol. 28. n° 11. 1420-1428. (2012)

[9]i.cs.hku.hk/~alse/hkubrg/projects/idba_ud/