Tratamiento de Secuencias de ADN y Clustering de Pacientes con Cáncer Colorrectal.

Laura Avila*, Victoria Santa María**, Luis López*, Marcelo Soria***, Cristóbal R. Santa María*,
*DIIT-UNLaM, **Instituto Lanari-FMed-UBA, ***FAUBA
Florencio Varela 1903 San Justo Pcia. de Buenos Aires
54-011-44808952
laura_avila75@yahoo.com.ar
vctrsntmr@hotmail.com
llopez@ing.unlam.edu.ar
soria@agro.uba.ar
csantamaria@unlam.edu.ar

RESUMEN

Con este trabajo se continua la línea de investigación consistente en evaluar y desarrollar procedimientos computacionales adecuados para analizar la relación clínica entre el microbioma intestinal y la presencia del cáncer colorrectal. En esta oportunidad se ha trabajado con muestras propias obtenidas en el medio local. Corresponden a los microbiomas de 20 pacientes, 10 sanos y 10 enfermos, del Sector de Coloproctología del Hospital Italiano de Buenos Aires que fueron secuenciados a partir de materia fecal. La identificación bacteriana se realizó utilizando el gen marcador 16S rRNA para obtener la distribución de frecuencias a distintos niveles taxonómicos en cada paciente. El presente artículo describe el proceso que se ha realizado desde que las muestras salen del secuenciador hasta que son procesadas para su valoración clínica. Con tal objetivo los pacientes fueron agrupados por medio de algoritmos de aprendizaje no supervisado y se desarrolló el aspecto matemático de una distancia que trata de ajustar el clustering computacional a los objetivos clínicos. La metodología de trabajo empleada ha sido validada mediante la participación en la red Global Research Network to Investigate the CRC-associated Microbiome of non-Western Countries creada por la Universidad de Leeds, UK.

Palabras Clave: Secuencias, Microbioma, Clasificación, Cáncer Colorrectal

CONTEXTO

El Grupo de Investigación y Desarrollo en Data Mining del Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLaM viene realizando evaluaciones y desarrollos de algoritmos con el fin de evidenciar los aspectos médicos de interés para diagnosticar y observar la evolución de patologías gastrointestinales tales como el cáncer colorrectal. Con tal finalidad desde 2015 ha desarrollado, dentro del programa de Incentivos, los proyectos de investigación C169 “Aplicaciones de Data Mining al Microbioma Humano” y C200 “Aplicación de Técnicas de Data Mining para Análisis del Microbioma Humano según Funcionalidades Metabólicas”. Actualmente lleva adelante, por primera vez a partir de muestras tomadas a pacientes autóctonos, el proyecto C220 del mismo Programa, “Explotación de Datos del Microbioma de Pacientes con Cáncer Colorrectal” en el marco de un convenio de colaboración con el Hospital Italiano de Buenos Aires firmado entre UNLAM e HIBA durante 2019.

1. INTRODUCCIÓN

Se estima que hasta el 90% de las condiciones de salud y enfermedad están asociadas de alguna manera al microbioma. Por ese motivo y por la posibilidad de intervenciones con prebióticos y antibióticos, los estudios metagenómicos basados en la Secuenciación de Nueva

Generación abren una nueva era en la prevención y tratamiento [1]. El cáncer colorrectal, que presenta características moleculares particulares y estrecha relación con la dieta "occidental" [2], es una patología de estudio de las más frecuentemente abordadas debido a su alta incidencia. La metagenómica orientada hacia el uso de genes marcadores como el 16S rRNA permite establecer el perfil taxonómico del microbioma de pacientes con cáncer colorrectal. En este camino es posible que en algún momento el análisis del microbioma alcance a transformarse en una herramienta auxiliar para el diagnóstico y evaluación de la enfermedad. Sin embargo, toda esta potencialidad depende en gran medida de que sea ajustada la interrelación entre lo bioinformático y lo médico. Cada algoritmo a utilizar, cada parámetro a ajustar, requieren de una evaluación acerca del grado en que colaboran a mejorar, en términos médicos, la herramienta de análisis. En tal sentido el vasto campo que constituye el dominio de las técnicas de proceso desde que se obtienen las secuencias de ADN del secuenciador hasta el desarrollo de métodos de aprendizaje automático que afinen la precisión en la clasificación médica, define la problemática a investigar por el grupo en el marco del convenio referido con el Hospital Italiano de Buenos Aires.

2. LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

El trabajo pretende estudiar en detalle la aplicación de procedimientos computacionales supervisados y no supervisados sobre los microbiomas para clasificar y predecir patologías. Comprende tanto el enfoque a través del gen marcador, el caso de los resultados que aquí se presentan, como el enfoque a partir de la información de funcionalidad metabólica aportada por el metagenoma completo. Se intentan alcanzar varios objetivos:

-Dominar la tecnología de almacenamiento, comparación y distribución funcional según las secuencias obtenidas del microbioma intestinal de pacientes por videocolonoscopía o por materia fecal.

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

-Realizar lo propio respecto de algoritmos de predicción entrenados y testeados para la evaluación clínica.

-Dejar allanado el camino para la aplicación experimental de todos estos métodos a mayor cantidad de muestras de pacientes locales obtenidas por investigadores del grupo.

El primer trabajo que ha sido necesario realizar fue el de establecer la secuencia de procesos para poder hacer el análisis estadístico posterior. En la Figura 1 se muestra el diagrama de los procesos que se han efectuado.

Importación de los reads. Para su importación a QIIME2 [3] y [4], se ha creado un archivo delimitado por comas (.csv), en formato *fastq manifest*. Este archivo es el que ha permitido conectar los identificadores de las muestras con las rutas absolutas de los archivos fastq.gz que contiene las secuencias directas e inversas, indicando la dirección de la lectura para cada una.

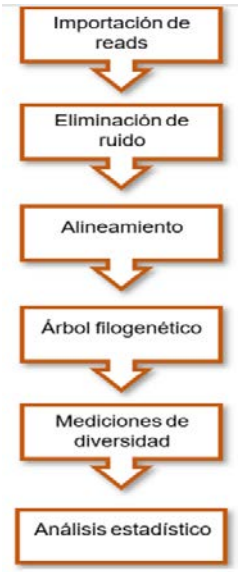

Figura 1. Pasos del proceso

Eliminación del ruido. En este paso del proceso se realiza el filtrado de las secuencias y se

eliminan las lecturas ambiguas o de baja calidad. También se realiza la eliminación de quimeras para evitar la falsa diversidad que se podría generar en análisis posteriores. Asimismo, se descartan las muestras cuyo recuento final sea inferior a 10000. Como resultado de dicha etapa se genera una tabla de frecuencias de las secuencias agrupadas en Unidades Taxonómicas Operacionales (OTU) como representantes de las lecturas.

Alineamiento. Se ha creado una tabla de secuencias alineadas, mediante algoritmos de alineamiento múltiple.

Árbol filogenético. Antes de realizar el árbol, ha sido necesario filtrar las secuencias, ya que el proceso de alineamiento añade ruido.

Mediciones de diversidad. Todo el trabajo realizado en los pasos previos ha permitido el estudio de la diversidad alfa y beta, mediante métricas filogenéticas y no filogenéticas [5]. Las medidas de alfa diversidad que se han utilizado son: Índice de diversidad de Shannon, OTUs observadas, análisis de correlación de Spearman. En cuanto a las medidas de diversidad beta, se han realizado análisis con diferentes distancias: Jaccard, Bray-Curtis, unweigthted Unifrac, weighted Unifrac y test de Adonis.

Se ha analizado la composición microbiana de las muestras según el grupo de pertenencia (sano o con presencia de cáncer de colon) y se ha aplicado test de Kruskal-Wallis a dichas matrices para determinar si existen diferencias estadísticamente significativas entre los grupos. Para ello se ha observado la relación entre la cantidad de OTUs y la diversidad de Shannon, y las variables categóricas de la metadata.

Para el análisis de diversidad beta se tuvieron en cuenta varias métricas y se realizó el análisis de permutaciones sobre las matrices de distancia (permutaciones) PERMANOVA, en la búsqueda de diferencias en la composición de los grupos de pertenencia con respecto al género. Por otro lado, se generaron gráficos a partir del análisis de componentes principales con las distintas métricas, que permitieron visualizar los resultados de la diversidad beta.

3. RESULTADOS OBTENIDOS/ESPERADOS

Los análisis de diversidad alfa realizados con las métricas de Shannon y por OTUs observadas, confirman que no existen sesgos en cuanto a su condición clínica o a su género. Como ejemplo, se presenta la Figura 2 y la Tabla 1.

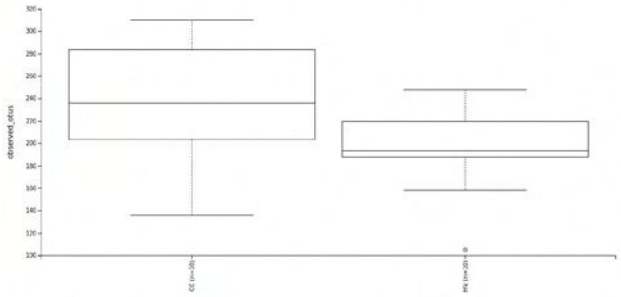

Figura 1. Gráfico de significación de diversidad de OTUs observadas dividida según el grupo (CC: cáncer de colon, HV: sano) al que pertenecen las muestras.

Resultados Kruskal-Wallis (todos los grupos)	
H	0,0701
p-valor	0,7911

Tabla 1. Test de Kruskal-Wallis para la diversidad de OTUs observadas según el grupo (CC: cáncer de colon, HV: sano) al que pertenecen las muestras.

Además de analizar la diversidad alfa, se analizó también la diversidad beta con diferentes métricas, puesto que este tipo de diversidad informa sobre el grado de diferenciación entre comunidades microbianas. Este análisis se ha realizado mediante componentes principales y usando el análisis de permutaciones PERMANOVA sobre las matrices de distancias. Ninguno de ellos reveló diferencias en la

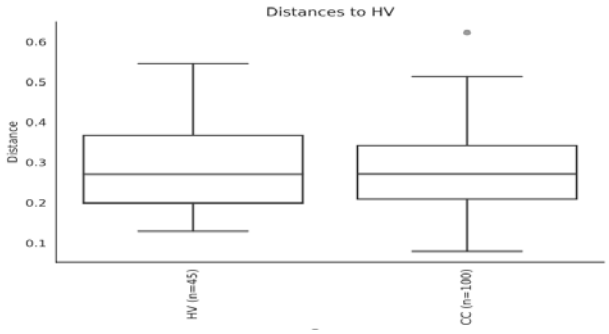

Figura 3. Gráfico que muestra la Unifrac entre cada grupo de interés respecto a los pacientes sanos.

composición entre los grupos, como se ve en la

Figura 3, que se muestra como ejemplo ya que los demás resultan similares.
El test estadístico aplicado para confirmar que las diferencias no son significativas ha sido el test PERMANOVA. Sus resultados se ven en la Tabla 2.

Resultados de PERMANOVA

Método	PERMANOVA
Nombre del test estadístico	Pseudo-F
Medida de la muestra	20
Número de grupos	2
Test estadístico	0,97616
p-valor	0,385
Número de permutaciones	999

Tabla 2. Test permanova a nivel taxonómico género.

Para abordar el análisis de la composición taxonómica de las muestras según los grupos de interés, en QIIME2 se ha asignado taxonomía a la tabla de secuencias representativas mediante el clasificador SILVA 132. Luego se colapsa la taxonomía a nivel 6 (nivel especie), en una tabla rarefaccionada, que permite exportarse para realizar otros tipos de análisis estadísticos.
La tabla de frecuencias taxonómicas rarefaccionada que se ha exportado del QIIME2 se cruzó con los metadatos de las muestras, es decir, la tabla obtenida incluye, además de las frecuencias absolutas de cada uno de los 239 Otus halladas, la clasificación en sano o enfermo, la edad y el sexo de cada paciente.
La información obtenida a la salida de QIIME2 se dispuso en tablas donde cada fila representa un microbioma, es decir un paciente, y en cada columna se ubican los taxones correspondientes al nivel taxonómico que se tiene en cuenta. Así hay tablas por género, familia, orden, clase y phylum que son los niveles a los cuales se realizan los estudios. La Tabla 3 muestra un ejemplo:

Columna1	OTU1	OTU2	OTU3	OTU4	OTU5	OTU6
GCRFNG_AR	0	0	5	0	42	0
GCRFNG_AR	0	0	0	0	22	0
GCRFNG_AR	160	0	2	0	213	0
GCRFNG_AR	0	0	0	0	359	1

Tabla 3. Ejemplo de tabulación a nivel género.

En la Tabla 3 la columna 1 corresponde a la identificación del paciente y las siguientes se asocian a las distintas Unidades Taxonómica Operacionales en las que se han agrupado las secuencias del gen marcador en cada microbioma. El número dentro de cada celda de la Tabla 3 es la cantidad de veces que la OTU respectiva se ha presentado en el correspondiente microbioma o, lo que es lo mismo, la cantidad de secuencias que han sido asignadas en ese microbioma a ese taxón. El total de OTUS halladas en todos los pacientes fue de 239. Es decir, se hallaron 239 géneros distintos en los que distribuir las secuencias de los genes marcadores aunque, claramente, no en todos los microbiomas se presentaron todos los géneros. La información de la tabla incluye en las tres últimas columnas, que no se ven, la clasificación en sano o enfermo, la edad y el sexo de cada paciente. A partir de ella se realizaron distintos procesos. El cálculo de la correlación lineal entre las variables y la clasificación de enfermedad o salud dio, como se esperaba, alta correlación lineal (-0.78) entre la edad y la enfermedad con un valor p del orden de 10^{-5} lo que autoriza a sostener tal correlación no solo en la muestra sino a nivel poblacional. A continuación, se realizó un clustering no jerárquico utilizando la distancia euclídea y encadenamiento promedio. Se tomaron en cuenta solo las variables que correspondían a cada taxón descartando la edad, el sexo y la clasificación médica respecto de la enfermedad. La Tabla 4 muestra los resultados.

Cluster	N° de Pacientes	Silueta
1	18	0,31
2	2	-0,14
Total	20	0,26

Tabla 4. Agrupamientos de pacientes

Es claro que el agrupamiento de bajo índice silueta total no revela nada sobre la condición clínica de los pacientes. La correlación entre la variable de clasificación de la enfermedad y la variable conglomerado asignado es exactamente

0 con un valor p de exactamente 1 lo que indica la imposibilidad de rechazar la hipótesis de no correlación a nivel poblacional. Sin embargo, se aprecian ciertas diferencias importantes de frecuencias promedio entre pacientes enfermos y sanos. Para evaluar la influencia real de cada diferencia en la disimilaridad de casos enfermos y sanos, las frecuencias medias pueden ser estandarizadas y luego calcular las diferencias para cada taxón (en el ejemplo el género u OTU). Tales diferencias se toman en valor absoluto mediante la cuenta $D_i = |fCC_i - fHV_i|$. Así se obtiene un perfil de diferencias de frecuencias medias estandarizadas entre pacientes sanos y enfermos como el que se muestra en la Figura 4 para el nivel taxonómico género.

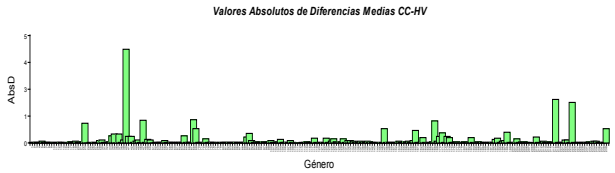

Figura 4. Diferencia de frecuencias medias sanos-enfermos

A continuación se realiza la cuenta $P_j = 1000 \frac{D_j}{\sum_{k=1}^{239} D_k}$ para todo j=1,2,…,239 obteniéndose un peso para cada diferencia. Con estos pesos se arma una distancia entre microbiomas i y k cuya fórmula es: $d = \sqrt{\sum_{j=1}^{239} (f_{ij} - f_{kj})^2 P_j}$. Esta distancia se propone para un nuevo clustering no jerárquico con encadenamiento promedio. Se observa entonces que los casos enfermos son todos bien clasificados, mientras que solo resultan bien clasificados la mitad de los pacientes sanos. La correlación lineal entre ambas variables arroja un coeficiente de 0.58 y el valor p fue 0.01 lo que indica que puede rechazarse la inexistencia de correlación con una probabilidad 0.01 de error. La nueva distancia pesada parece desempeñarse mejor para evaluar la similitud entre pacientes de acuerdo a su clasificación clínica. Los pesos obtenidos podrían muy bien utilizarse para medir distancias entre casos que no hayan integrado la muestra original. Resulta auspicioso que todos los casos enfermos hayan sido bien clasificados, pues la correlación de Spearman, que se utiliza también en variables cualitativas, dio relativamente alta y con muy baja probabilidad de error al extenderse a la población.

4. FORMACIÓN DE RECURSOS HUMANOS

En el equipo de trabajo participan un magister y un especialista en data mining, un doctor en biología, un médico, 2 ingenieros en sistemas y una matemática Está en curso una tesis de maestría.

5. BIBLIOGRAFÍA

[1] Di Bella, et al. 2013. High throughput sequencing methods and analysis for microbiome research. Journal of Microbiological Methods. Vol. 95, Issue 3, pp 401-414. https://doi.org/10.1016/j.mimet.2013.08.011

[2] Carbonetto, B., et al. 2016. Human Microbiota of the Argentine Population- A Pilot Study. Frontiers in microbiology, 7, 51. https://doi.org/10.3389/fmicb.2016.00051

[3] Bolyen E, et al. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857. https://doi.org/10.1038/s41587-019-0209-9

[4] D'Argenio V, et al. 2014. Comparative Metagenomic Analysis of Human Gut Microbiome Composition Using Two Different Bioinformatic Pipelines. BioMed Research International. Vol. 2014 Article ID 325340 https://doi.org/10.1155/2014/325340

[5] Xia, Y., & Sun, J. (2017). Hypothesis Testing and Statistical Analysis of Microbiome. Genes & diseases, 4(3), 138–148. https://doi.org/10.1016/j.gendis.2017.06.001