Minería de Datos para Análisis del Microbioma Humano

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

Resumen

Se expone la línea de investigación que lleva adelante el Grupo de Investigación y Desarrollo en Data Mining del Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLaM. Se detallan los resultados del proyecto de investigación "Aplicaciones de Data Mining al Estudio del Microbioma Humano", C169 del Programa de Incentivos. La línea de trabajo intenta aportar procedimientos computacionales adecuados para analizar la relación clínica entre el microbioma intestinal y la presencia de patologías tales como el cáncer de colon y la enfermedad de Crohn. El trabajo hasta aquí realizado comprende la obtención de una muestra de microbiomas de pacientes desde el repositorio de NCBI, la identificación bacteriana a partir del gen marcador y la determinación de la distribución de frecuencias por especies en cada paciente. Se continuó luego con el agrupamiento de pacientes por enterotipos y la evaluación clínica de las categorías obtenidas.

Palabras Clave: Microbioma Secuencias Clasificación Diagnóstico

Contexto

Desde los comienzos de su vida el cuerpo humano es colonizado por bacterias, arqueas, hongos y virus. Esta comunidad de microorganismos se denomina microbioma y contiene diez veces más células que las del propio cuerpo humano. La cantidad de genes presentes en total es varios órdenes de magnitud mayor que la del genoma humano. Si bien esto se conoce desde hace largo tiempo, la imposibilidad de cultivo en laboratorio de la mayoría de esos microorganismos ha dificultado hasta ahora análisis profundos. La nueva generación de tecnologías de secuenciación de ADN ha permitido comenzar a estudiar las características del microbioma humano según la edad de la persona, la localización geográfica, los hábitos alimentarios y la presencia de enfermedades. El objetivo principal de estos estudios metagenómicos es analizar la estructura y la dinámica de comunidades microbianas, para establecer cómo se relacionan sus miembros entre sí, cuáles son las sustancias que producen y que consumen, y especialmente cuálesson sus interacciones con las células humanas próximas y cómo se modifica la comunidad en presencia de enfermedades. El estudio por medio de la identificación de un gen marcador como el 16s rRNA pretende evaluar características ecológicas como la riqueza y la diversidad, mientras que el análisis del metagenoma como un todo, identificando las secuencias obtenidas por comparación con una base

de datos genética previamente armada, permite agrupar los genes identificados por funciones metabólicas asociadas con la presencia de enfermedades. Los estudios que emplean tales técnicas se multiplican velozmente y existen a nivel mundial proyectos de investigación como el Metagenomics of the Human Intestinal Tract (MetaHIT) o el Human Microbiome Proyect (HMP). En nuestro país en el marco del Plan Nacional de Ciencia, Tecnología e Innovación (Argentina Innovadora 2020), dentro del área de Salud, se ha comenzado a desarrollar una Plataforma Tecnológica de Genómica y Bioinformática que facilitará estudios similares. El objetivo general es entender el funcionamiento del microbioma humano a partir del procesamiento y análisis de muestras de secuencias de ADN, y desarrollar nuevas herramientas para analizar y caracterizar el curso de patologías poniendo énfasis en el cáncer de colon y en la enfermedad de Crohn.

Introducción

El trabajo computacional consiste en obtener los datos secuenciados de una muestra integrada por varios microbiomas. Cada microbioma debe cotejarse con una base de datos correspondiente a un gen marcador para encontrar la distribución de frecuencias de los microorganismos identificados por tal gen. Alternativamente el conjunto de secuencias del microbioma puede 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 [2]. En cualquier caso 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 a una especie o género distinto, y el valor que adopta es la cantidad presente del respectivo microorganismo o grupo (taxón). Armados con las distancias y los encadenamientos adecuados, los diferentes clusters constituyen los enterotipos. 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

La línea de trabajo pretende estudiar en detalle la aplicación de métodos 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. Se intentan alcanzar varios objetivos:

-Dominar la tecnología de almacenamiento, comparación y distribución funcional según las secuencias obtenidas del microbioma intestinal de pacientes por vía preferente de videocolonoscopía o alternativamente 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 muestras de pacientes locales obtenidas por investigadores del grupo.
-En tal sentido, obtener muestras de pacientes en el medio local, analizar las características del protocolo médico a seguir, enviarlas a secuenciar y aplicar los procedimientos ya probados sobre muestras no propias.

Resultados y Objetivos

Un primer aspecto a resolver fue el del hardware y sistema operativo necesario para el trabajo. Con vistas al desarrollo completo de la línea de investigación dentro del Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLAM se decidió adaptar un servidor con ocho procesadores Intel-I7 y 16 GB de memoria RAM para el trabajo del grupo, ya que el volumen de memoria y la capacidad de proceso requeridas por los conjuntos de datos utilizados en la bibliografía y los que potencialmente pudieran formarse luego a partir de muestras propias, lo hacían necesario. También se decidió dotar al servidor con el sistema BioLinux, de uso habitual para este tipo de trabajos pues posee una fácil interacción con paquetes de software libre, eficientes y probados, en la investigación en biología computacional.
Se trabajó con una muestra de 11 pacientes, 7 sanos y 4 afectados por la enfermedad de Crohn [3], y para realizar las asignaciones taxonómicas y funcionales se utilizó el software de reciente desarrollo, SUPERFOCUS, que efectúa ambas tareas. Para la asignación funcional usa la base de datos SEED [4]. Las salidas de SUPERFOCUS son varios archivos de texto, uno con las asignaciones funcionales, tres con la información de subsistemas y otro más con la información taxonómica. Para consolidar la información obtenida de todos los bloques se programaron dos scripts en R para procesar la información taxonómica y la funcional. Ambos scripts leen los archivos de una muestra, agregan la información y producen como salida dos tipos de matrices. Para la información taxonómica se producen una serie de matrices con la asignación taxonómica en una dimensión y la muestra en la otra. Existen tablas para cada una de las categorías taxonómicas usadas en biología: reino, phylum, clase, orden, familia, género y especie. En el caso de la información funcional, la salida está constituida por tablas para la función específica y tres más para cada uno de los subsistemas que define el proyecto SEED.
Con el archivo de asignaciones taxonómicas se inició el proceso utilizando el software R para armar la matriz de distancias. Para medir la distancia entre dos microbiomas cuyas distribuciones de de frecuencias estadísticas de especies son conocidas hay que medir las distancias entre ambas distribuciones. Para ello se utiliza la distancia de Jensen-Shanon [5] Siendo P y Q dos distribuciones conocidas puede definirse $R = \frac{1}{2}(P + Q)$ y su entropía $\kappa = H(R)$. Luego de algunas consideraciones [5] resulta $H(R) \approx \frac{1}{2}H(P) + \frac{1}{2}H(Q)$ y se define la expresión

$$D_P^2 = 2H(R) - H(P) - H(Q)$$

Entonces

$$\sqrt{D_P^2} = \left(2H(R) - H(P) - H(Q)\right)^{\frac{1}{2}}$$

cumple con las condiciones matemáticas para ser una distancia entre P y Q [10].
De acuerdo a ello queda

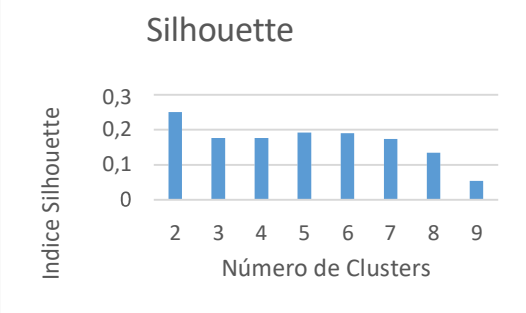
$$D_F = \sum_{i=1}^{N} p_i \log \quad p_i + q_i \log q_i)^{\frac{1}{2}}$$

como expresión de la distancia entre dos distribuciones de frecuencias.
Esta distancia es la que usa el algoritmo PAM (Partitioning around medoids) utilizado en este caso directamente de la

biblioteca cluster de R. El medoide es el elemento para el cual la disimilitud promedio con todos los objetos en el conglomerado es mínima.
En [2] se utiliza para definir el número óptimo de clusters, el índice de Calinski-Harabasz [6] que se construye de la siguiente forma:

$$C = \frac{\frac{B_k}{k-1}}{\frac{W_k}{n-k}}$$

Aquí B_k es la suma de las distancias al cuadrado de todos los elementos *i* y *j* que no pertenecen al mismo cluster, W_k es la suma de los cuadrados de las distancias de todos los elementos *i* y *j* que pertenecen al mismo cluster, *n* es el número de elementos a clasificar y *k* la cantidad seleccionada de clusters. Sin embargo dada la cantidad exigua de microbiomas disponibles en la muestra, en el presente estudio se evaluó solo el agrupamiento obtenido con el índice Silhouette para lo cual se realizaron los agrupamientos considerando distintos números de clusters. Los resultados pueden apreciarse en la Figura 1.
Figura 1.



Se decidió entonces considerar el agrupamiento en 2 enterotipos y realizar sobre él una reducción de las 277 dimensiones de la muestra a través de un análisis de componentes principales. Se vio entonces que las primeras dos componentes explican el 53% y el 21% respectivamente de la información.

El análisis de correlación lineal de estas dos componentes CP1 y CP2 con las variables originales arrojó los siguientes datos significativos volcados respectivamente en las Tablas 1 y 2
Tabla 1

CP1(53%)	Alta Correlación
[Ruminococcus]_obeum	-0,92
Bacteroides_helcogenes	0,89
Bacteroides_salanitronis	0,83
Bacteroides_thetaiotaomicr..	0,92
Bacteroides_vulgatus	0,97
Bacteroides_xylanisolvens	0,96
Kitasatospora_setae	0,81
Porphyromonas_gingivalis	0,96
Prevotella_denticola	0,85

Tabla 2

CP2(21%)	Alta Correlación
butyrate-producing_bacteri..	0,75
Clostridium_saccharolyticu..	0,75
Roseburia_hominis	-0,75
Roseburia_intestinalis	-0,79

Las Tablas 1 y 2 permiten ver que el 74% de la información aportada por el agrupamiento realizado se explica predominantemente por la presencia de las bacterias citadas que tienen alta correlación con las componentes principales 1 y 2. Si bien desde el punto de vista clínico este conocimiento puede representar un aporte importante no bastó, en este caso para que los dos enterotipos obtenidos se correspondieran con la clasificación de sano o enfermo previamente determinada por otros procedimientos diagnósticos. La Figura 3 muestra los enterotipos obtenidos. El Enterotipo 1 es coloreado en azul mientras que el Enterotipo 2 es amarillo. En la Figura 4 se ven los mismos agrupamientos pero ahora la diferencia de color entre los casos señala la presencia o ausencia de la enfermedad.

Figura 3

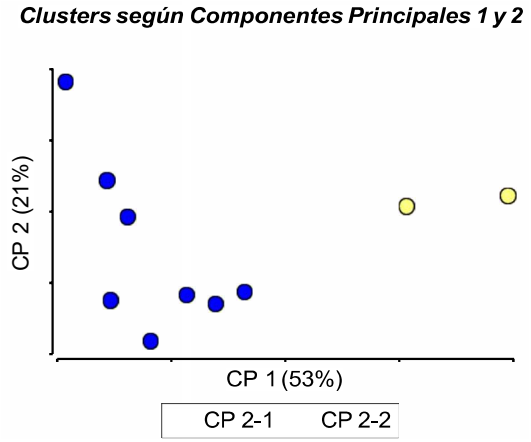
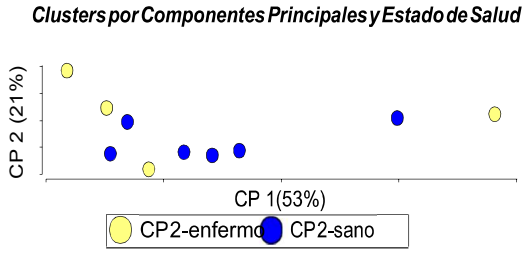


Figura 4



Los resultados obtenidos confirman que si bien la utilización del gen marcador no bastó para lograr una clasificación adecuada, en 3 de los 4 casos parece confirmarse la manifestación de la Enfermedad de Crohn asociada con la baja presencia de las especies Bacteroides y Clostridium que aparecen en alta correlación directa con los bajos valores de las CP1 y CP2. [7] Se concluye entonces en la necesidad de ampliar el trabajo por vía del estudio metagenómico de biomarcadores funcionales. Pararealizarlo en vez hallar los enterotipos por clasificación taxonómica habrá que establecerlos por grupos ortólogos (OG) que codifican para distintas enzimas o proteínas que pertenecen a distintas vías metabólicas. [6]

Formación de Recursos Humanos

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

Referencias

[1] Ngom-Bru, Catherine and Barretto, Caroline. Gut microbiota: methodological aspects to describe taxonomy and functionality. Briefings in Informatics. Vol3 NO 6. 747-750. 2012

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

[3] Morgan, XC. et al. (2012) Dysfunction of the intestinal microbiome in inflammatory bowel disease and treatment. Genome Biology 2012, 13:R79

[4] http://theseed.org

[5] Endres, D y Schindeling,J. A New Metric for Probability Distributions. IEEE. Transactions on Information Theory. Vol. 49 NO.7. 2003.

[6] Calinski, T., and J. Harabasz. "A dendrite method for cluster analysis." Communications in Statistics. Vol. 3, No. 1, 1974, pp. 1–27.

[7] Ray K. IBD. Understanding gut microbiota in new-onset Crohn's disease. Nat Rev Gastroenterol Hepatol [Internet]. Nature Publishing Group;2014;11(5):268.