# Minería de Datos del Microbioma en Pacientes con Cáncer Colorectal

*Cristóbal Santa María UNLaM csantamaria@unla.edu.ar*
*Laura Ávila UNLaMLaura_Avila75@yahoo.com.ar*
*Victoria Santa María FM-UBA vctrstmr@gmail.com*
*Luis López UNLAM llopezar@yahoo.com.ar*
Marcelo Soria FCEyN soria@agro.uba.ar

## Resumen

*Los métodos de secuenciación de ADN permiten estudiar comunidades enteras de microorganismos tal como la que constituye el microbioma intestinal humano. Hay un creciente interés médico en este análisis pues las modificaciones que ocurren en la microbiota pueden ser responsables de la disbiosis asociada con enfermedades como el cáncer colorectal sobre el cual se focaliza este trabajo. El proceso bioinformático de las muestras desde que salen las lecturas del secuenciador hasta que pueden ser explotadas como datos, requiere sistematizar y automatizar las tareas estableciendo una "pipeline" de procesos ligados a través de programas confeccionados al efecto. También es necesario validar el uso de algoritmos de aprendizaje utilizados en la propia explotación de los datos, para hallar asociaciones y patrones que vinculan la clasificación taxonómica y las vías metabólicas presentes con la condición clínica de los pacientes analizados. El presente trabajo encara ambas tareas. Sobre muestras de lecturas "crudas" de microbiomas colónicos, obtenidas del repositorio internacional NCBI (National Center of Biotechnology Information), primero describe y realiza la secuencia de procesos hasta que se obtiene la identificación taxonómica a nivel del taxón Phylum y las anotaciones funcionales KO. A continuación, elije y evalúa algoritmos de clustering jerárquico, realiza un análisis de componentes principales y estudia la aplicación de árboles de decisión y ensambles sobre esos conjuntos de datos. Se logra entonces poner a punto el proceso de secuencias de ADN obtenidas al utilizar secuenciadores de nueva generación sobre todo el metagenoma. De esta forma se ha podido encarar una segunda etapa de análisis, actualmente en desarrollo, con muestras extraídas de pacientes autóctonos. Estos estudios, procuran avanzar y establecer en nuestro medio, la caracterización clínica del cáncer colorectal por esta vía tecnológica.*

## Introducción

Las tecnologías de nueva generación para la secuenciación de ADN permiten tratar cada vez mayor cantidad de muestras a menores costos. Esto ha potenciado notablemente las posibilidades de los estudios metagenómicos, que involucran el conocimiento simultáneo de los genes de todos los individuos que forman una comunidad, extendiendo sus alcances al análisis de la composición microbiana de suelos, aguas y al microbioma humano. Éste no es otra cosa que la comunidad de microorganismos presentes en el cuerpo humano que contiene diez veces más microorganismos que células propias. Se han presentado entonces probabilidades ciertas de evaluar la interacción entre esta microbiota y el organismo alojante que resulta clave en el mantenimiento de la inmunidad y la protección contra agentes patógenos externos al organismo humano. La composición del microbioma varía entre las personas según el estilo de vida, la dieta y su genotipo, pero es estable dentro de una misma persona. Si se producen modificaciones de tipo permanente esto conlleva una disbiosis que es la alteración de la influencia de la comunidad en los procesos metabólicos de la persona y que se asocia con enfermedades tales como la inflamación intestinal, el asma o los desórdenes mentales. En particular la disbiosis está implicada en la carcinogénesis al ser iniciadora de los procesos inflamatorios y su presencia da señal de inmunodepresión [1]

Algunos argumentos indirectos sugieren el rol potencial de la microbiota intestinal en la carcinogénesis colorectal. El cáncer colorectal es básicamente una enfermedad genética pero el microbioma alojado por el paciente puede explicar la interacción entre los genes del paciente y el entorno de microorganismos presentes que se manifiesta tanto en su diversidad y riqueza taxonómica cuanto en las vías metabólicas que tienen lugar. Frecuentemente parecen asociados el cáncer colorectal y las variaciones de las frecuencias con que algunas especies bacterianas se encuentran en el microbioma [2] Esta asociación no es clara aún para determinar si la variación del microbioma es una causa o un efecto del cáncer. Incluso recientemente se ha sugerido que el microbioma puede jugar el rol de control sobre la enfermedad. En todo caso existe una perspectiva interesante en los estudios metagenómicos pues no solo permiten la determinación taxonómica de la comunidad microbiana a través de la utilización de genes marcadores sino que también, al utilizar la información de todos las secuencias obtenidas del microbioma (WGS), pueden establecer las vías metabólicas que potencialmente sigan los procesos celulares en el paciente. Esto ha motivado un profundo interés en la

comunidad médica que ha buscado así relacionarse con los campos de la biología, la computación, la estadística y específicamente con la bioinformática para avanzar en la comprensión y eventualmente en el diagnóstico y pronóstico de enfermedades.

Al respecto de lo hasta aquí expuesto debe señalarse que no solo los avances en la tecnología de secuenciación han permitido estos estudios cada vez más amplios y profundos, sino que también se han producido importantes desarrollos de algoritmos cuya rapidez y precisión de cálculo ha sido creciente a la vez que ha permitido una mayor cantidad de procesos de explotación de datos tanto en aspectos estadísticos descriptivos cuanto en técnicas de aprendizaje automático supervisado y no supervisado. En lo referido al microbioma humano se ha hecho evidente la necesidad de contar con un esquema seriado de procesos computacionales a aplicar desde que las secuencias salen del secuenciador hasta que resultan transformadas en información útil para la investigación clínica. Esto involucra la confección de software de filtrado de las secuencias, de evaluación de contaminación del conjunto con secuencias humanas, de ensamblado de secuencias, de anotación de las mismas según sus niveles taxonómicos, de identificación de vías metabólicas presentes, de agrupamiento en conglomerados o clusters según taxonomía o metabolismo, de aprendizaje sobre conjuntos de entrenamiento y testeo para clasificar microbiomas según los mismos principios. Una gran mayoría de estos desarrollos se realizan en forma de software libre para que puedan ser utilizados y testeados por investigadores a nivel global y se encuentran muchas veces disponibles en repositorios internacionales que también contienen los datos de las distintas experiencias realizadas.

El trabajo consistió entonces en la construcción de una "pipeline" que permitiese llegar desde las secuencias crudas hasta el proceso bioinformático de aprendizaje automático para clasificación de casos. Se tomó un conjunto de secuencias de microbiomas correspondiente a pacientes con antecedentes de cáncer de colon extraído del repositorio de NCBI [3] y se fueron realizando en línea los distintos procesos que pudieran finalmente revelar aspectos clínicos de interés. Se analizaron así aspectos matemáticos del tratamiento masivo de datos de ADN metagenómico y se desarrollaron programas en lenguaje R o C para lograr unir cada etapa del proceso con la siguiente. Se procuró además constatar los propios resultados con los obtenidos en estudios similares. También se comenzó a trabajar en la obtención de una muestra de pacientes locales a efecto de caracterizar en el futuro similitudes y diferencias clínicas al aplicarle la línea de procesos ahora establecida.

## Materiales y Métodos

Se inició el análisis de las muestras correspondientes a secuenciación de ADN total de un estudio depositado en la base de datos BIOPROJECT del NCBI con el código PRJNA397450. Este estudio consistió en la extracción y secuenciación de ADN microbiano total y del gen que codifica para el ARN ribosomal de 16S (16S rRNA), a partir de muestras de hisopados rectales, biopsias por endoscopia y materia fecal. Seleccionados dentro de un programa de prevención del cáncer colorectal según reglas estándar tanto estadísticas como de protocolo médico, los pacientes elegidos fueron adultos de entre 40 y 85 años, de buena salud, con antecedentes de pólipos colorectales. Se tomaron muestras en dos momentos diferentes de tiempo espaciados en tres meses. Se procuró además que en la muestra total hubiera un 50% de pacientes con recurrencia en los pólipos y un 50% que no. Finalmente 60 individuos fueron seleccionados a lo largo de dos años sobre un total de 150 participantes. Más características de la muestra utilizada se explicitan en [4].

Los datos de secuenciación del estudio son entonces públicos y al momento de comenzar el trabajo no tenían ninguna publicación asociada. Esto no fue un inconveniente, pues lo necesario era que los datos permitieran poner a punto, y en lo posible automatizar, las operaciones bioinformáticas necesarias: la selección de software más adecuado, el análisis de calidad de las secuencias, el diseño de una metodología de limpieza de las secuencias, su posterior validación, el ensamblado de los metagenomas y finalmente la implementación de la "pipeline" que permitiera integrar todos los pasos y automatizar la mayor cantidad. 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. La 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 que se detallan a continuación se realizaron en computadoras corriendo el sistema operativo Linux, distribución Ubuntu 16.04. Se procedió a la descarga de los 286 archivos utilizando la herramienta SRATOOLKIT que distribuye el 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 de estas secuencias con el software FastQC [5] y se determinó que casi todas las secuencias tenían restos de dos de los adaptadores que usa Illumina para la secuenciación, una en las secuencias F ("forward") y otro en las secuencias 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 en que ocurría se utilizó el software Scythe [6]. El proceso de limpieza se realizó con el programa Cutadapt [7] que permite realizar limpieza de adaptadores, cortes por caída

en los valores de calidad, eliminación por largo mínimos, cortes en posiciones arbitrarias, etc. En primer término, se realizaron una serie de pruebas preliminares para determinar las opciones específicas de limpieza y los valores óptimos de los diferentes 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 resultó con una longitud menor a 50 bases se procedió 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 aquellos con al menos un miembro de longitud menor a 50 bases. Después de este proceso de limpieza se volvió a revisar la calidad de las secuencias con FastQC, que mostró resultados satisfactorios.

Con el objetivo de obtener la caracterización taxonómica y funcional de las muestras se llevaron a cabo luego distintos procesos. Por un lado, se trabajó directamente con los reads. que son las lecturas obtenidas de la secuenciación, y por otro se enfocó la tarea hacia la obtención de los contigs (reads ensamblados). Se procuró establecer adecuadamente la cadena de procesos [8] en este último caso. La pipeline utilizada fue la siguiente:

### 1. Ensamblado.

El proceso de ensamblado consiste en unir los reads para obtener secuencias más largas, contigs, que se corresponden con las secuencias de ADN antes de haber sido cortadas para su secuenciación. Al respecto se estudiaron en detalle los aspectos matemáticos [9] y se realizaron pruebas con los programas IDBA-UD [10] y Megahit [11]. Finalmente se decidió utilizar Megahit, con una longitud mínima de 400 bases por contig. [12]

### 2. Identificación y eliminación de secuencias humanas.

Una vez obtenidos los contigs, se eliminaron aquellos correspondientes a la contaminación humana. Para este proceso se alinearon los contigs contra el genoma humano de referencia GRCh38, utilizando Blastn. Los contigs que alinearon contra el genoma humano se eliminaron del juego de datos utilizando un script R desarrollado para tal fin.

### 3. Anotación taxonómica.

La anotación taxonómica de los contigs no humanos se realizó utilizando Kaiju junto con su propia base de datos. [13] y [14]

### 4. Anotación funcional.

La anotación funcional inicial de los contigs no humanos se realizó con Prokka. Luego se agregaron anotaciones KEGG usando el servicio KAAS [15]

### 5. Proceso de Recuento:

A continuación, se usó el software Bowtie2 [16] para generar los índices de los contigs y luego se alinearon los reads contra esos contigs. De esta forma se obtuvieron los mapeos de los reads a la referencia en archivos sam. Luego, utilizando samtools se los convirtió a formato bam y se indexaron estos mapas. Con samtools se obtuvo la cantidad de reads que mapearon abriendo la perspectiva a estudios del microbioma basados directamente en los reads, aunque se prefirió por ahora analizarlo desde los contigs pues se poseía al respecto mayor información.

### 6. Consolidación de los resultados

Para llevar a cabo los procesos de análisis estadístico y minería de datos se construyeron una serie de scripts en R que consolidan en una única tabla la información de anotación funcional para cada gen predicho con Prokka y KAAS, de anotación taxónomica de KAIJU a nivel de contigs y variables adicionales como el largo de cada contig, el largo de cada gen predicho. La forma en que allí se disponen los datos puede verse en las líneas de ejemplo de la Tabla 1

**Tabla 1. Ejemplo de información consolidada**

| sample_ID | contig_ID | ID | reino | phylum | KO |
|---|---|---|---|---|---|
| SRR5907479 | k141_3 | GKLBDHMD_00001 | Bacteria | Actinobacteria | NA |
| SRR5907479 | k141_4 | GKLBDHMD_00002 | Bacteria | Firmicutes | K03282 |
| SRR5907479 | k141_5 | NA | Bacteria | Firmicutes | NA |
| SRR5907479 | k141_7 | GKLBDHMD_00003 | NA | NA | K01424 |

En la Tabla 1 la columna primera identifica la muestra, la segunda el contig y luego, de reino a género, se anotan los distintos niveles taxonómicos de la rama del árbol filogenético a la que el contig pertenece. La columna KO contiene las anotaciones de las posibles funcionalidades metabólicas de la secuencia según la Kyoto Encyclopedia of Genes and Genomes (KEGG) [17] lo que permitirá estudiar cada microbioma y al conjunto de todos los que conforman la muestra desde el punto de vista de las funcionalidades metabólicas que puedan estar presentes.

### 7. Disposición de los datos para su explotación en tablas de frecuencias

El trabajo realizado hasta aquí lleva a las puertas del procesamiento estadístico inteligente. Sin embargo, resta aún la importante tarea de disponer los datos en esquemas y formatos que puedan ser leídos y procesados por los paquetes estadísticos y de aprendizaje automático. Tal tarea se resolvió desarrollando dos programas en lenguaje C que devolvieron las frecuencias absolutas con que los distintos taxones y anotaciones KO se presentaban en cada microbioma. De esta forma se eligió trabajar con las tablas Microbiomas.Phylum y Microbiomas.KO pues la primera da información sobre los Phylum, entre ellos por ejemplo las fusobacterias cuya especie Fusobacteruim Nucleatum resulta prevalente en el carcinoma humano colorectal [2]. Así mismo las anotaciones KO permitirían clasificar los microbiomas, es decir los pacientes, según las funciones metabólicas que pudieran estar teniendo lugar a nivel celular, identificando con ello condiciones clínicas de interés. Las Tablas 2 y 3 muestran un fragmento a título de ejemplo de la forma en que se dispone la información sobre

frecuencias estadísticas para los Phylum y las anotaciones KO respectivamente.

**Tabla 2. Ejemplo de frecuencias estadísticas para Phylum**

| Microbioma | Acidobacteria | Actinobacteria | Aquificae | Armatimonadetes | Bacteroidetes |
|---|---|---|---|---|---|
| PhSRR5907479 | 0 | 108 | 0 | 1 | 49 |
| PhSRR5907480 | 29 | 935 | 10 | 0 | 1046 |
| PhSRR5907481 | 16 | 394 | 1 | 0 | 570 |
| PhSRR5907482 | 10 | 507 | 1 | 2 | 778 |
| PhSRR5907483 | 34 | 770 | 4 | 4 | 627 |
| PhSRR5907484 | 1 | 102 | 0 | 0 | 55 |
| PhSRR5907485 | 41 | 1443 | 9 | 4 | 1926 |
| PhSRR5907486 | 16 | 779 | 4 | 3 | 664 |
| PhSRR5907487 | 13 | 197 | 1 | 1 | 139 |
| PhSRR5907488 | 3 | 126 | 0 | 0 | 84 |
| PhSRR5907490 | 7 | 313 | 1 | 4 | 309 |
| PhSRR5907491 | 2 | 118 | 0 | 0 | 56 |
| PhSRR5907492 | 0 | 3 | 0 | 0 | 1 |

En la primera columna de la Tabla 2 se ven los códigos relativos a cada microbioma. Desde la segunda columna en adelante hasta la número 55 corresponden a los nombres de los distintos Phylum que fueron hallados entre las muestras. Cada número en el interior de la tabla corresponde a la frecuencia absoluta de un taxón Phylum en un dado microbioma. Lo propio ocurre con la Tabla 3 donde las columnas son las anotaciones KO halladas que en este caso fueron 4000.

**Tabla 3. Ejemplo de frecuencias estadísticas para anotaciones KO**

| Microbioma | K00002 | K00003 | K00005 | K00008 | K00009 | K00010 | K00011 | K00012 |
|---|---|---|---|---|---|---|---|---|
| SRR5907479 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SRR5907480 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 |
| SRR5907481 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SRR5907482 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SRR5907483 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 |
| SRR5907484 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| SRR5907485 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 3 |

## 8. Explotación de datos

Para procesar los datos de las tablas Microbiomas.Phylum y Microbiomas.KO se utilizaron los paquetes Weka [18] e Infostat [19]. Los procedimientos aplicados sobre los datos a nivel taxonómico Phylum fueron los siguientes:

- Para obtener un perfil inicial sobre la diversidad y riqueza de los distintos Phylum en la muestra total, se calcularon medidas estadísticas de resumen determinando para cada Phylum su media y suma total.

- A continuación, se procuró determinar alguna forma de agrupamiento que mostrara posibles diferencias que analizadas luego clínicamente tuvieran relevancia. Se analizaron distintas metodologías para formar conglomerados. Para establecer los clusters en forma jerárquica se utilizó la distancia euclídea con encadenamiento promedio eligiéndose la formación de 2 agrupamientos que se volcaron en el dendrograma de la Figura 1.

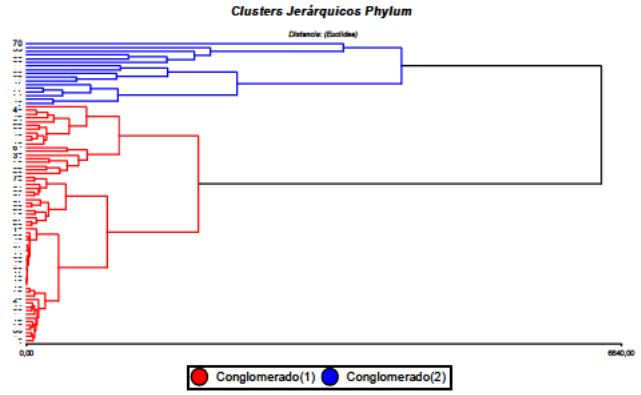


**Figura 1. Clusters Jerárquicos Phylum**

- En la búsqueda de diferencias entre los conglomerados, se obtuvieron los respectivos perfiles promedio que se muestran en las Figuras 2 y 3.

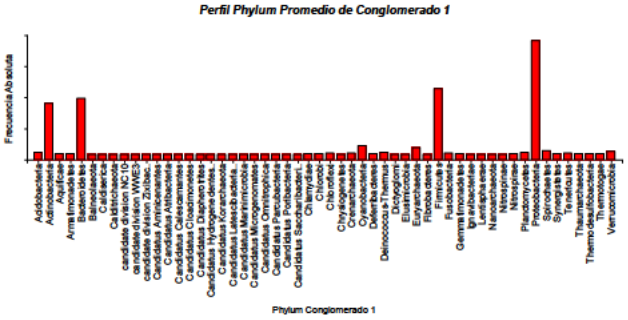


**Figura 2. Perfil Phylum de Promedios-Conglomerado 1**

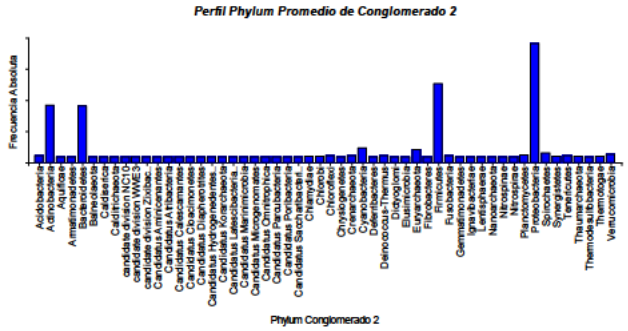


**Figura 3. Perfil Phylum de Promedios-Conglomerado 2**

- Luego, a efecto de reducir la cantidad de variables buscando explicar al menos la parte substancial de la clasificación por conglomerado, se aplicó sobre todo el conjunto de microbiomas, el método de componentes principales. Para todo el conjunto de microbiomas se seleccionaron las dos primeras variables: CP1, que explica el 71% del comportamiento de las variables Phylum, y CP2 que explica el 4%.

- Para determinar aquellos Phylum mas relacionados estadísticamente entre si y a su vez los más relacionados con la variable de clasificación conglomerado, se estudió la correlación lineal entre variables. También se analizó la relevancia de cada variable Phylum en la determinación del conglomerado por vía de un algoritmo de selección de atributos que evalúa el valor de cada subconjunto de atributos considerando la habilidad

predictiva en cada caso junto al grado de redundancia entre las variables [20]. Los resultados se ven en la Tabla 4.

**Tabla 4. Selección de atributos**

| Ranking | Phylum |
|---|---|
| 1 | Calditrichaeota |
| 2 | candidate division NC10 |
| 3 | candidate division Zixibac.. |
| 4 | Candidatus Aminicenantes |
| 5 | Candidatus Calescamantes |
| 6 | Candidatus Saccharibacteri |
| 7 | Chlamddiae |
| 8 | Fibrobacteres |
| 9 | Fusobacteria |
| 10 | Nanoarchaeota |

- A continuación se consideraron por separado los conglomerados y se calcularon las componentes principales con el objetivo de estudiar en cada caso la representatividad que pudieran ostentar sobre las variables Phylum. Se estableció que para el Conglomerado 1 el eje CP1 explica el 49% del comportamiento de la variable Phylum mientras que el CP2 agrega un 5% más. A su vez en el Conglomerado 2 la componente principal CP1 representa el 48% del comportamiento de Phylum pero el CP2 alcanza un 11%. Además, para cada conglomerado se calculó la correlación de Phylum con sus componentes principales

- Finalmente se utilizó la variable Conglomerado a efecto de entrenar y testear un árbol de decisión para predecir la clasificación de un paciente aún no catalogado en ninguno de los conglomerados. Primero se realizó el entrenamiento de un árbol J48 mediante un subconjunto formado por 38 microbiomas. Luego se llevó a cabo el testeo con un subconjunto distinto de 43 microbiomas. La Tabla 5 muestra el desempeño del árbol ya en su fase de testeo.

**Tabla 5. Testeo árbolJ48 phylum**

| | | |
|---|---|---|
| Correctly Classified Instances | 39 | 90.6977 % |
| Incorrectly Classified Instances | 4 | 9.3023 % |

| TP Rate | FP Rate | ROC Area | Class |
|---|---|---|---|
| 0,889 | 0,088 | 0,900 | b |
| 0,912 | 0,111 | 0,900 | a |

=== Confusion Matrix ===

| a | b | <-- classified as |
|---|---|---|
| 8 | 1 | a = b |
| 3 | 31 | b = a |

De acuerdo a la evaluación de Stantikov et al [21] el algoritmo Random Forest [22] resulta el de óptimo desempeño en datos de microbiomas por lo que se procedió a construir el ensamble de árboles respectivo. En este caso los resultados obtenidos en la fase de testeo se muestran en la Tabla 6.

**Tabla 6. Testeo ensamble Random Forest phylum**

| | | |
|---|---|---|
| Correctly Classified Instances | 43 | 100 % |
| Incorrectly Classified Instances | 0 | 0 % |

| TP Rate | FP Rate | ROC Area | Class |
|---|---|---|---|
| 1,000 | 0,000 | 1,000 | b |
| 1,000 | 0,000 | 1,000 | a |

=== Confusion Matrix ===

| a | b | <-- classified as |
|---|---|---|
| 9 | 0 | a = b |
| 0 | 34 | b = a |

Sobre las anotaciones KO asociadas con funcionalidades metabólicas se realizaron los procesos que se detallan a continuación:

-Se calcularon medidas estadísticas de resumen determinando para cada anotación KO su suma, máximo y promedio sobre todo el conjunto de microbiomas.

-Se formaron conglomerados por el método jerárquico incorporándose la nueva variable Conglomerado con valores 1 o 2 a la tabla de Microbiomas.KO. La Figura 4 muestra el dendrograma correspondiente.

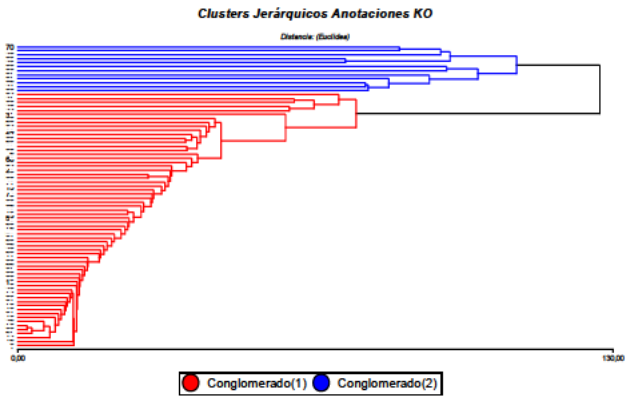


**Figura 4. Clusters Jerárquicos Anotaciones KO**

- Se estudió la influencia de cada variable en la asignación del conglomerado respectivo a través del algoritmo de selección de atributos ya citado.

- El conjunto de microbiomas se particionó para obtener un conjunto de entrenamiento y otro de testeo a fin de establecer un árbol de decisión que permita clasificar cualquier microbioma, aún no estudiado, dentro uno de los conglomerados establecidos. Las Tabla 7 muestra el desempeño del ensamble Random Forest en fase de testeo.

**Tabla 7. Testeo ensamble Random Forest anotaciones KO**

| | | |
|---|---|---|
| Correctly Classified Instances | 35 | 94.5946 % |
| Incorrectly Classified Instances | 2 | 5.4054 % |

| TP Rate | FP Rate | ROC Area | Class |
|---|---|---|---|
| 1,000 | 0,333 | 0,995 | a |
| 0,667 | 0,000 | 0,995 | b |

=== Confusion Matrix ===

| a | b | <-- classified as |
|---|---|---|
| 31 | 0 | a = a |
| 2 | 4 | b = b |

### 9. Datos de pacientes autóctonos

A partir de un convenio firmado entre la UNLAM y el Hospital Italiano de Buenos Aires el Sector de Coloproctología ha extraído 20 muestras de materia fecal, 10 de pacientes on cáncer colorectal y otras 10 de personas sanas que forman parte del material que actualmente se procesa y analiza.

## Resultados

El primer resultado importante es el establecimiento de una pipeline de procesos a realizar, con el conocimiento acabado de las técnicas y programas involucrados, con la confección de distintos programas que unen las diferentes partes y con la experiencia acerca de los resultados a obtener cuando se la aplica. Este es un punto significativo pues por lo general las distintas metodologías empleadas o no se revelan claramente en los artículos o se lo hace fragmentariamente de modo tal que resultaba imperioso conocerlo a efecto del análisis de muestras propias. A su vez, este conocimiento obtenido permitirá introducir mejoras en los procesos estudiando con mayor detalle sus variantes posibles tanto desde el punto de vista algorítmico como del de programación. La Figura 5 da cuenta, a modo de resumen, del flujo que deben seguir los datos desde que salen desde el secuenciador hasta que están en condiciones de prestarse a interpretación médica.

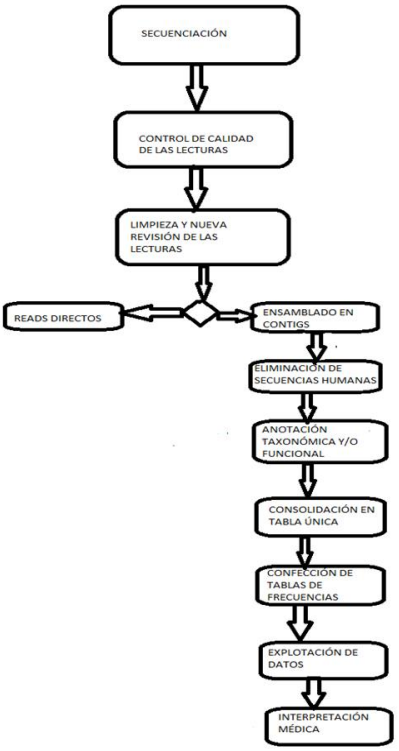


**Figura 5. Flujo de Datos**

Corresponde hacer algunos comentarios sobre la explotación de los datos realizada. A nivel taxonómico Phylum, se detectaron 4 categorías de gran abundancia relativa en los pacientes: Actinobacteria, Bacteroidetes, Firmicutes y Proteobacteria. A su vez con abundancia relativa media se encontraron Cianobacteria y Euriarcheota. El resto presentó una baja abundancia relativa, incluso el Phylum Fusobacteria cuya especie Fusobacteria Nucleatum es prevalente en el cáncer colorectal. Los dos conglomerados obtenidos a partir del conjunto de microbiomas no revelan diferencias esenciales en los promedios de frecuencia por Phylum respecto del perfil total como tampoco lo hicieron para las otras medidas de resumen que fueron consideradas. Se realizó el análisis de componentes principales con el objetivo de reducir las variables y se encontró que sobre el plano de ambas componentes principales los microbiomas del primer conglomerado están más concentrados mientras que, los del segundo conglomerado están mucho más dispersos. Además, a la luz de los perfiles de correlación obtenidos con estas componentes principales, se estaría mostrando que los microbiomas del segundo conglomerado resultan más abundantes en términos ecológicos a nivel Phylum. En la búsqueda de hallar causas a las diferencias en la asignación de conglomerados se estudió la correlación entre las variables Phylum y la Conglomerado. La más baja correlación con ésta correspondió a Canditaus

Diapherotrites lo que está pendiente de interpretación clínica.

Así se continuó con el modelado por medio de árboles de decisión de la clasificación microbiómica a nivel Phylum. Se probó en primer término un árbol J48 y luego, a efecto de mejorar el desempeño, con el modelo de ensamble de árboles denominado Random Forest que al ser testeado tuvo un porcentaje del 100% de casos bien clasificados. (Tabla 6)

También se analizaron los microbiomas según las funcionalidades KO. Esta información permite comprender la vía metabólica que a nivel celular puede estar presente y en términos médicos asociar ese proceso con la condición clínica del paciente. También en el caso de las Anotaciones KO se obtuvieron dos conglomerados por el método jerárquico (Figura 4). Finalmente se entrenó un ensamble random forest cuyo desempeño resultó muy bueno. Para el conjunto de testeo el modelo arrojó un porcentaje de casos bien clasificados del 94.6% como puede verse en la Tabla 7.

## Discusión

Tal cual se recoge en la vasta y actual bibliografía sobre el tema, la metagenómica puede colaborar en el estudio del cáncer colorectal ayudando a desentrañar la etiología de algunos de sus procesos al buscar la caracterización adecuada del microbioma, su riqueza y diversidad. También es posible que en algún momento alcance a transformarse en una herramienta auxiliar al diagnóstico y a la evaluación del estadío 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 este estudio se ha tenido cuidado en no aplicar programas como una caja negra donde entran datos y sale información. Se lo ha hecho así en la certeza de que la interpretación correcta de una información requiere algún conocimiento conceptual acerca de cómo se ha obtenido. Por supuesto se trata de un campo interdisciplinario en el cual las preguntas y las respuestas pueden venir de distintos especialistas. En este caso se ha visto que en la formación computacional de los conglomerados intervienen muy distintos aspectos que pueden modificar su composición significativamente. El primero es si se debe utilizar un método jerárquico o uno como k-means. A nivel Phylum se observó que resultaba la misma clasificación. En segundo término, para formar los dos clusters se utilizó el encadenamiento promedio. Una tarea a efectuar es probar con los distintos tipos de encadenamiento posible. La tercera cuestión es la forma de medir la distancia. Se usó aquí la común distancia euclídea pero habría que probar la efectividad de la distancia entre distribuciones de probabilidad dado que son frecuencias absolutas o relativas los valores que las distintas variables adoptan [23]. En cada caso es el mayor o menor poder de revelar la información de tipo médico, el que debe guiar la elección que entonces ya no depende del criterio estadístico o informático solamente. Otro asunto importante lo constituye la preparación de un paquete unificado de programas, escritos en un mismo tipo de código. Aquí varios de los programas necesarios fueron confeccionados en lenguaje R, otros en cambio se escribieron en C, también se ensayó algo con el lenguaje Python y, por supuesto se aplicó una serie de programas ya realizados de código abierto o en versiones liberadas. Cada especialista conoce software relativo a su ejercicio profesional, pero automatizar toda la tarea involucrada en la pipeline desarrollada sería más fácil si se la unifica en alguna forma encadenando computacionalmente los procesos que pueden ir invocándose en serie.

## Conclusiones

Se ha podido establecer una pipeline con varios pasos automatizados para tratar las secuencias de ADN microbiómico desde que salen del secuenciador hasta que resultan datos para explotación. Por técnicas estadísticas multivariadas y de aprendizaje no supervisado, como el clustering,, pueden formarse grupos que permitan caracterizaciones clínicas. Esta clasificación puede ser utilizada predictivamente mediante el entrenamiento y testeo de árboles de decisión que, como se ha comprobado, muestran un desempeño óptimo. El trabajo debe continuar afinando la interpretación médica sobre muestras de pacientes propios, tarea en la cual ya se está trabajando para lograr un marco más amplio de validación de los resultados.

## Referencias

[1] Lopez, A et al. “Microbiota in digestive cancers: our new partner?” Carcinogenesis, 2017, pp. 1-10, doi:10.1093/carcin/bgx087

[2] Castellarin, M et al. “Fusobacterium nucleatum infection is prevalent in human colorectal carcinoma”, Genome Research.2012, pp. 299-306.

[3] https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA397450

[4] Jones, R B. et al. “Inter-niche and inter-individual variation in gut microbial community assessment using stool, rectal swab and mucosal samples”. Scientific Reports. Volume 8, 2018, Article number: 4139. www.nature.com/scientificreports

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

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

[7] https://cutadapt.readthedocs.io/en/stable/

[8]Coil, D; Jospin, G; and Darling, A. “A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data”, Bioinformatics, Vol. 31, n° 4, 2015, pp. 587-589.

[9] Santa María, C; Rebrij, R; Santa María, V and Soria, M. “Treatment of Massive Metagenomic Data with Graphs” en Libro de Actas, VI Jornadas de Cloud Computing & Big Data. La Plata, Argentina, Junio 27-29, 2018, pp 77-80

[10] 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, 2012, pp. 1420-1428.

[11] Li, D; Liu, CM; Luo, R; Sadakane, K and Lam, TW. “MEGAHIT: an ultra-fast single-node solution for large and

complex metagenomics assembly via succinct de Bruijn graph", Bioinformatics, Vol. 31, n°10, 2015, pp 1674-1676 https://academic.oup.com/bioinformatics/article/31/10/1674/177884)

[12] https://github.com/voutcn/megahit

[13] http://kaiju.binf.ku.dk/

[14] https://www.nature.com/articles/ncomms11257.

[15] https://www.genome.jp/kegg/kaas/

[16] http://bowtie-bio.sourceforge.net/bowtie2/ y

https://www.nature.com/articles/nmeth.1923

[17] M. Kanehisa, Y. Sato; M. Furumichi , K. Morishima y M. Tanabe. "New approach for understanding genome variations in KEGG". Nucleic Acids Research, Volume 47, Issue D1, 8 January 2019, pp. D590–D595, https://doi.org/10.1093/nar/gky962

[18] https://www.cs.waikato.ac.nz/ml/weka/

[19] http://www.infostat.com.ar/

[20] M. A. Hall. Correlation-based Feature Subset Selection for Machine Learning. Hamilton. New Zeland. 1988

[21] Statnikov A. Henaff M. Narendra V. Konganti K. Li Z. Yang L. Pei Z. Blaser M. AliferisC y Alekseyenko A. "A comprehensive evaluation of multicategory classification methods for microbiomic data", Microbiome, 2013, pp. 1:11

[22] Breiman, Leo (2001). "Random Forests". Machine Learning 45 (1), 2001, pp. 5–32. doi:10.1023/A:1010933404324.

[23] Endres, D y Schindeling,J. "A New Metric for Probability Distributions". IEEE.Transactions on Information Theory, Vol.49, n° 7, 2003.