Departamento: Ingeniería e Investigaciones Tecnológicas
Programa de acreditación:
PROINCE

Programa de Investigación[1]:

Código del Proyecto: C220

Título del proyecto
Explotación de Datos del Microbioma de Pacientes con Cáncer Colorectal

Director:
Santa María, Cristóbal Raúl

Codirector:
López, Luis

Integrantes:
Ávila, Laura Cacho Mendoza, Ariel Martínez, Pablo Otaegui, Juan Carlos
Investigador Externo
Soria, Marcelo
Asesor- Especialista
Santa María, Victoria

Resolución Rectoral de acreditación: N° 348/2019

Fecha de inicio:
01/01/2019

Fecha de finalización:
31/12/2020

[1] Los Programas de Investigación de la UNLaM están acreditados con resolución rectoral, según lo indica la Resolución HCS Nº 014/15 sobre **Lineamientos generales para el establecimiento, desarrollo y gestión de Programas de Investigación a desarrollarse en la Universidad Nacional de La Matanza**. Consultar en el departamento académico correspondiente la inscripción del proyecto en un Programa acreditado.

A. Desarrollo del proyecto

A.1. Grado de ejecución de los objetivos inicialmente planteados, modificaciones o ampliaciones u obstáculos encontrados para su realización (desarrolle en no más de dos (2) páginas)

En lo referido al desarrollo del proyecto de investigación deben citarse dos hechos que lo condicionaron significativamente, a pesar de los cuales pudo cumplirse con lo esencial del trabajo previsto. En primer lugar, recién pudo contarse con muestras de pacientes locales tomadas en el servicio de coloproctología del HIBA en junio de 2019. En segundo término, durante marzo de 2020, en razón de la pandemia, la toma de nuevas muestras se interrumpió sin reanudarse hasta la fecha de finalización pautada para el proyecto. Ambos hechos limitaron la cantidad de pacientes que integraron los estudios por lo que los resultados obtenidos deben considerarse iniciales. Por otra parte, las condiciones de trabajo de todo el equipo durante el año 2020 se redujeron sensiblemente teniendo que apelar a la virtualidad para la comunicación y también para el uso de recursos y equipos. Aún dentro de este esquema de funcionamiento se considera haber logrado resultados de interés si bien tampoco pudo ejecutarse el presupuesto asignado en forma completa habida cuenta que, ante la falta de nuevas muestras, no hubo necesidad de invertir en parte de los insumos previstos para procesarlas. Se detalla a continuación el trabajo realizado.

A efecto de hacer concordar las tareas con las llevadas adelante en el Hospital Italiano de Buenos Aires y en la Universidad de Leeds del Reino Unido, y poder así validar la metodología de trabajo, se utilizó el gen marcador 16S rRNA. Se ha trabajado con dos muestras de materia fecal de pacientes del Servicio de Gastroenterología, sector de Coloproctología del HIBA. La primera, Muestra 1, estuvo integrada por 20 pacientes de los cuales 10 presentaban cáncer de colon ya diagnosticado y otros 10 estaban sanos. La segunda muestra, Muestra 2, fue recogida sobre 15 pacientes, 7 de ellos con cáncer colorrectal y 8 sanos.

La secuenciación de la primera muestra se realizó en un secuenciador Illumina HiSeq sobre la región V4 del gen 16S rRNA, obteniéndose 150 pares de bases por cada secuencia. Con el mismo equipamiento se realizó la secuenciación de la segunda muestra, sobre las regiones V3 y V4 del gen 16S rRNA, lo que arrojó lecturas de 300 pares de bases.

Para procesar las lecturas se ha utilizado el software QIIME2, que es una plataforma bioinformática que permite trabajar con secuencias obtenidas de secuenciadores de nueva generación. QIIME2 es de código abierto, extensible, gratuita y desarrollada por la comunidad. Proporciona herramientas de visualización interactiva para realizar análisis exploratorio y mostrar informes de resultados.

Descripción de los procesos realizados.

1- El proceso validado

a-Importación de las secuencias. QIIME requiere que los archivos se importen en formato comprimido y con sus respectivos puntajes de calidad. Las secuencias proporcionadas en la primera muestra se encuentran en dicho formato, por lo que solo es necesario importarlas a una carpeta propia del programa. Los datos que contienen son, 20 archivos

comprimidos, uno por cada paciente, con los reads en forma directa, 20 archivos comprimidos, también de cada paciente, con los reads en forma inversa, un archivo que contiene las rutas de acceso a cada uno de los 40 archivos mencionados antes y un archivo de metadatos que contiene la información que han recolectado en el hospital de cada paciente, es decir, condición clínica, sexo y edad, entre otros. Para su importación a QIIME2, 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. Con la segunda muestra de 15 pacientes se procedió en forma enteramente análoga.

b-Eliminación de ruido. Para realizar el proceso de filtrado de ambas muestras se ha usado el plugin *dada2*, en QIIME2, con el método *denoise-paired.* En este paso del proceso se realiza el
filtrado de las secuencias y se eliminan las lecturas ambiguas o de baja calidad, teniendo en cuenta que se eliminan todas aquellas que tienen un puntaje de calidad inferior a 19, lo que implica una probabilidad de error de la base inferior a 0,012. También se realiza la eliminación de quimeras. Las quimeras son fragmentos de secuencias que se producen por la técnica de secuenciación, para evitar la falsa diversidad que se podrían generar en análisis posteriores. Como resultado de dicha etapa se genera una tabla de frecuencias de las secuencias por cada una de las dos muestras consideradas, agrupadas en Unidades Taxonómicas Operacionales (OTU) como representantes de las lecturas. El umbral de agrupamiento es del 97% de identidad.

c-Alineamiento. Para poder determinar si los conjuntos de secuencias de los pacientes diferían en su composición de acuerdo a su grupo de pertenencia (sano o con presencia de cáncer de colon), o si diferían en riqueza o abundancia, se ha desarrolla el árbol filogenético, para poder obtener medidas de diversidad filogenéticas. En principio se realizó el alineamiento empleando el *plugin alignment* con el método *mafft*, para realizar múltiples alineamientos de las secuencias representativas. Como resultado se ha crea una tabla de secuencias alineadas, que se utiliza para realizar el árbol.

d-Árbol filogenético. Antes de realizar el árbol, se filtraron las secuencias, ya que el proceso de alineamiento añade ruido; para ello se utilizó el *plugin mask.* Finalmente se generó el árbol filogenético con esas secuencias filtradas, utilizando el *plugin phylogeny* con el método *fasttree,* Como el método produce un árbol sin raíz, se le aplica el método *midpoint-root,* que produce el enraizamiento del punto medio para colocar la raíz del árbol en el punto medio de la distancia de punta a punta más larga en el árbol sin raíz.

e-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, con el *plugin diversity* empleando la pipeline *core-metrics-phylogenetic.* Usando este método se ha podido rarefaccionar la tabla de OTUs a una profundidad de 36259 (determinada durante la visualización de la tabla de frecuencias). Es necesario realizar este paso ya que las métricas de diversidad son sensibles a la profundidad de las muestras. En este proceso se extraen secuencias al azar sin reemplazo de cada muestra y todas terminan teniendo la

misma profundidad. Las medidas de alfa diversidad que se han realizado son: Índice de diversidad de Shannon, OTUs observadas, análisis de correlación de Speaman. En cuanto a las medidas de diversidad beta, se han realizado análisis con diferentes distancias: distancia de Jaccard, distancia de Bray-Curtis, distancia *unweigthted Unifrac*, distancia *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) con el *plugin diversity*, usando el método *diversity alpha-group-significance* sobre las tablas de medidas de diversidad alfa obtenidas 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. Luego se realizaron las curvas de rarefacción alfa, las cuales representan las medidas de diversidad en función de la profundidad del muestreo, que permiten comprobar si ha resultado adecuado el nivel de profundidad de secuenciación. Se ha utilizado el método *diversity alpha-correlation* para determinar si existe correlación entre la cantidad de OTUs observadas (como medida de riqueza) y la edad, y también entre la diversidad de Shannon y la edad.

Para el análisis de diversidad beta se ha utilizado el método *beta-group-significance* dentro del *plugin diversity*, teniendo en cuenta varias métricas y realizando 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.

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 obtenida con *dada2*, mediante el clasificador *silva-132-99-16S* dentro del plugin *feature-classifier* con el método *classify-consensus-blast*. Luego se colapsa la taxonomía a un nivel deseado con el *plugin taxa* y usando el método *collapse*, en este caso a nivel 5 (nivel género), a 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 muestra la frecuencia absoluta de cada género encontrada en cada microbioma o paciente. Esta tabla se cruzó con los metadatos para realizar en WEKA e INFOSTAT otros análisis relativos al entrenamiento de árboles de decisión.

Todos estos procesos se realizaron siguiendo la metodología estándar aplicada igualmente por el HIBA y por la Global Research Network to Investigate the CRC-associated Microbiome of non-Western Countries creada por la Universidad de Leeds, UK.

2- Clustering

Se realizaron distintos experimentos de agrupamiento de pacientes a efecto de evaluar las posibilidades de la técnica en la clasificación clínica adecuada de los pacientes de

acuerdo a su perfil microbiómico. Se realizaron pruebas de clustering jerárquico, con distancia euclídea, otras con agrupamiento no jerárquico por medio del algoritmo k-means, con distancia euclídea y encadenamiento promedio, variando el número inicial de centroides. Y finalmente se construyó "ad hoc" una distancia entre microbiomas que tiene en cuenta el peso de la diferencia de cada taxón entre pacientes sanos y enfermos.

Para evaluar la influencia real de cada diferencia en la disimilaridad de casos enfermos (CC) y sanos (HV), las frecuencias medias fueron estandarizadas y luego se calcularon las diferencias para cada taxón. Tales diferencias se tomaron en valor absoluto mediante la cuenta $D_i = |fCC_i - fHV_i|$. Así se obtuvo un perfil de diferencias de frecuencias medias estandarizadas entre pacientes sanos y enfermos para el nivel taxonómico género.

A continuación se realizó la cuenta $P_j = 1000 \frac{D_j}{\sum_{k=1}^{239} D_k}$ obteniéndose un peso para cada diferencia. Con estos pesos se armó 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 empleó para un nuevo clustering no jerárquico con encadenamiento promedio.

En estos procesos se utilizó software INFOSTAT, WEKA y desarrollos propios en lenguaje C para operar entre paquetes cambiando formatos y armar la matriz de distancias pesadas. Los agrupamientos fueron evaluados por el índice Silhouette

3- Árboles de decisión

En relación con los métodos de aprendizaje automático, en base a los antecedentes de desempeño, se decidió entrenar y testear dos algoritmos de árboles de decisión. Por un lado, el C4.5 disponible en Weka bajo el nombre J48 y por otro, el ensamble Random Forest, también incorporado a WEKA. Desde el punto de vista computacional se utilizaron matrices de confusión y curvas ROC

para evaluar tanto el entrenamiento, como el testeo. La consideración comparativa de las Muestras 1 y 2, requirió la identificación de los taxones presentes simultáneamente en ambas. Se identificaron
así 216 géneros comunes con los cuales se trabajó en los dos tipos de árboles. Además, los mismos algoritmos se probaron con la mezcla de Muestras 1y 2. Así se seleccionó convenientemente un conjunto de entrenamiento de 18 pacientes y otro de testeo de 17. En todos los casos, se estableció como criterio relevante en términos clínicos que la clasificación fuera muy eficiente en la detección de pacientes enfermos y menos importante en cuanto a la verificación de los sanos.

4- Metagenoma completo

En lo referente a los análisis sobre el metagenoma completo, en 2019 se efectuó además el ordenamiento de la pipeline, cuya metodología de trabajo difiere de la necesaria para el gen marcador. Las muestras utilizadas pertenecían a repositorios internacionales tales como NCBI y las distintas pruebas se habían realizado en 2018 dentro del proyecto del que el presente es continuidad. También estos aspectos requirieron algún ajuste al tratamiento informático:

a- un programa que toma como entrada los archivos generados para cada individuo en planilla de cálculo y que para cada caso del nivel taxonómico tiene distintas frecuencias. El programa genera un único archivo en el que cada fila representa un paciente y cada columna contiene las frecuencias del
caso al nivel taxonómico analizado. Esta aplicación, resuelta en Lenguaje C demandó unas 450 líneas de código.

b-Un programa similar al anterior que cambia las denominaciones según los niveles taxonómicos. Unas 90 líneas de código también en Lenguaje C.

c- Dado que las salidas de los programas citados son archivos de texto, y que para su posterior procesamiento en el 'pipeline' se requiere que los archivos estén en formato .xls o .xlsx, se generó un nuevo programa que genera en este formato archivos .csv. Se requirieron poco más de 200 líneas de código en Lenguaje C y VBS (script en visual basic).

d- Ante la necesidad de agrupar información correspondiente a Reino / Phylum / Clase / Orden / Familia / Género de diversas muestras en grupos con menor nivel de detalle (Reino/.../Famila - Reino/.../Orden – etc.), se diseñó un nuevo programa que lee el archivo de texto que genera una primera salida en que se acumulan los valores correspondientes a Género, una segunda salida en que se acumulan los valores correspondientes a Familia y Género, etc. Estas salidas están generadas como trasposición de la matriz original. A partir de estas múltiples salidas en formato .csv se producen los correspondientes en formato .xls Se requirieron 534 líneas de código en Lenguaje C.

e- Como se definieron distancias ad hoc para procesar las Unidades Taxonómicas Operacionales obtenidas a partir de las muestras de los pacientes autóctonos, se desarrolló una aplicación para calcularlas y generar la matriz cuadrada de distancias a cualquier nivel taxonómico. Se requirieron 846 líneas de código en lenguaje C.

Todo el desarrollo del proyecto se volcó pormenorizadamente en las publicaciones y presentaciones a congresos que se detallan y adjuntan.

B. Principales resultados de la investigación

B.3. Capítulos de libros

Autores	*Ávila, Laura | Santa María, Victoria | López, Luis | Soria, Marcelo | Santa María, Cristóbal*
Título del Capitulo	Base de Datos y Minería de datos
Título del Libro	*Actas del XXII Workshop de Investigadores en Ciencias de la Computación: WICC 2020*
Año	*2020*
Editores del libro/Compiladores	*Universidad Nacional de la Patagonia Austral- Red UNCI*
Lugar de impresión	
Arbitraje	*SI*
ISBN:	*978-987-3714-82-5*
URL de descarga del capítulo	*http://sedici.unlp.edu.ar/handle/10915/103591*
N° DOI	

Autores	*Ávila, Laura | Santa María, Victoria | López, Luis | Santa María, Cristóbal/ Marcelo Soria*
Título del Capitulo	Workshop Base de Datos y Minería de datos
Título del Libro	*Actas del XXVI Congreso Argentino de Ciencias de la Computación. CACIC 2020*
Año	*2020*
Editores del libro/Compiladores	*Universidad Nacional de La Lanza-Red UNCI*

Lugar de impresión	
Arbitraje	*SI*
ISBN:	*978-987-4417-90-9*
URL de descarga del capítulo	*https://ca-cic2020.un-lam.edu.ar/es-ar/pdf/2020-CA-CIC-LIBROACTAS-3.pdf*
N° DOI	

B.4. Trabajos presentados a congresos y/o seminarios

Autores	*Cristóbal Santa María, Laura Ávila, Victoria Santa María, Luis López y Marcelo Soria*
Título	*Minería de datos del microbioma en pacientes con cáncer colo-rectal*
Año	*2019*
Evento	*CONAIISI*
Lugar de realización	*San Justo- Buenos Aires*
Fecha de presentación de la ponencia	*14/11/2019*
Entidad que organiza	*UNLAM*
URL de descarga del trabajo (especificar solo si es la descarga del trabajo; formatos pdf, e-pub, etc.)	*https://conaiisi2019.un-lam.edu.ar/pdf/2019-CONAIISI-ACTAS-7MA-EDICION.pdf*

Autores	*Cristóbal Santa María*
Título	*Clustering y árboles de decisión en pacientes con cáncer colorectal*
Año	*2020*
Evento	*IV Encuentro del Programa MEP -Mejora de las Estrategias Pedagógicas*
Lugar de realización	*San Justo- Buenos Aires*
Fecha de presentación de la ponencia	*15/12/2020*
Entidad que organiza	*UNLAM*
URL de descarga del trabajo (especificar solo si es la descarga del trabajo; formatos pdf, e-pub, etc.)	

B.5. Otras publicaciones

Autores	Agencia CTyS-Unlam
Año	2020
Título	Analizan el vínculo entre el microbioma y el Cáncer de Colon
Medio de Publicación	http://www.ctys.com.ar/index.php?idPage=20&idArticulo=3697

Autores (de la parte realizada en Argentina-HI- UNLAM)	Cristóbal Santa María, Laura Ávila, Victoria Santa María, Luis López y Sector de Coloproctología del HIBA.

Año	2019
Título	The creation of a global research network to investigate the CRC-associated microbiome of non-Western countries
Medio de Publicación	Network Meeting. University of Leeds

C. Otros resultados. Indicar aquellos resultados pasibles de ser protegidos a través de instrumentos de propiedad intelectual, como patentes, derechos de autor, derechos de obtentor, etc. y desarrollos que no pueden ser protegidos por instrumentos de propiedad intelectual, como las tecnologías organizacionales y otros. Complete un cuadro por cada uno de estos dos tipos de productos.

C.2. Otros desarrollos no pasibles de ser protegidos por títulos de propiedad intelectual. Indicar: Producto y Descripción.

Producto	Descripción
Software en lenguaje C	opera entre distinto software empleado cambiando formatos y armando la matriz de distancias pesadas

D. Formación de recursos humanos. Trabajos finales de graduación, tesis de grado y posgrado. Completar un cuadro por cada uno de los trabajos generados en el marco del proyecto.

D.2. Tesis de posgrado: Maestría

Director (apellido y nombre)	Tesista (apellido y nombre)	Institución	Calificación	Fecha /En curso	Título de la tesis
Santa María, Cristóbal	Ávila, Laura	UNLAM		En Curso	Modelización para el Agrupamiento de Pacientes con Cáncer Colo-rectal

F. Vinculación: Indicar conformación de redes, intercambio científico, etc. con otros grupos de investigación; con el ámbito productivo o con entidades públicas. Desarrolle en no más de dos (2) páginas.

A instancias del desarrollo del proyecto el 20 de marzo de 2019 se firmó un convenio de cooperación y asistencia entre la UNLAM y el Institutos Universitario del Hospital Italiano IUHI. Esto permitió que el Sector de Coloproctología del Hospital Italiano de Buenos Aires aportara las muestras de pacientes autóctonos con las que se trabajó. En ese marco también se asistió a las I y II JORNADAS DE INVESTIGACIÓN TRASLACIONAL DE MICROBIOMA- DEL LABORATORIO A LA CLÍNICA organizadas por el IUHI. Por otra parte, a través de la IUHI y el Sector de Coloproctología se tomó contacto con la Large Bowel Microbiome Disease Network de la Universidad de Leeds, Inglaterra, lo que permitió comparar la obtención y el procesamiento bioinformático de las muestras propias y realizar los ajustes metodológicos a efectos de validar la pipeline desarrollada.

H. Cuerpo de anexos:

Anexo I:

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.

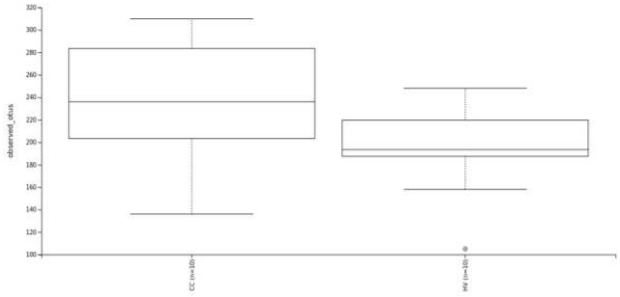


Figura 1. Gráfico de significación de diversidad de OTUs observadas dividida según el grupo (CC: cáncer de colon, HV: sano) al que pertenecen las muestras.

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

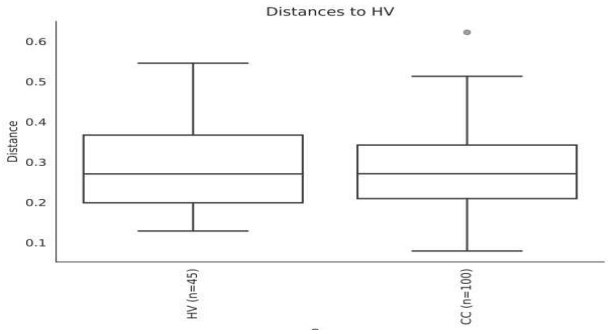


Figura 3. Gráfico que muestra la Unifrac entre cada grupo de interés respecto a los pacientes sanos.

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_AF	0	0	5	0	42	0
GCRFNG_AF	0	0	0	0	22	0
GCRFNG_AF	160	0	2	0	213	0
GCRFNG_AF	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.

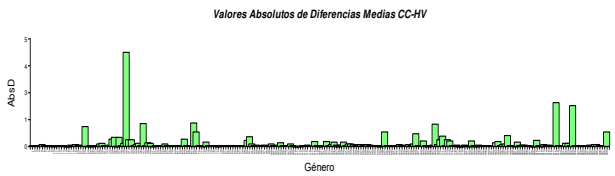


Figura 4. Diferencia de frecuencias medias sanos-enfermos

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

Se certifica que **CRISTÓBAL RAÚL SANTA MARIA (UNLAM)** ha participado en calidad de autor del artículo **TRATAMIENTO DE SECUENCIAS DE ADN Y CLUSTERING DE PACIENTES CON CÁNCER COLORRECTAL (12700 - BDMD)** aceptado en el XXII WORKSHOP DE INVESTIGADORES EN CIENCIAS DE LA COMPUTACIÓN – WICC 2020, organizado por la Universidad Nacional de la Patagonia Austral - Junio 2020.

Lic. Patricia Pesado
Coordinadora
RedUNCi

Ing. Hugo Santos ROJAS
Rector
UNPA

Análisis del Desempeño de Clustering y Árboles de Decisión en la Evaluación Clínica de Microbiomas 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

Abstract. 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. Cabe entonces explorar el papel del análisis taxonómico del microbioma como herramienta de diagnóstico y evaluación de la enfermedad. En tal sentido debe ajustarse la interrelación bioinformático-médica. Cada algoritmo a utilizar, cada parámetro a ajustar, requieren de una evaluación acerca del grado en que colaboran a mejorar el análisis en términos médicos. El objetivo general del trabajo es entonces caracterizar el microbioma de pacientes del AMBA en cuanto a riqueza, diversidad y distribución estadística, a través de muestras del gen marcador 16S rRNA obtenidas de materia fecal. En particular, se procuró reproducir la pipeline desarrollada anteriormente con muestras extraídas de repositorios internacionales mejorando los aspectos de automatización y ajustando la elección de parámetros. También se validó la metodología de trabajo por medio de comparación con los procesos llevados a cabo en el marco de la Large Bowel Microbiome Disease Network. A su vez, se realizó el análisis estadístico correspondiente para establecer la riqueza, diversidad de los microbiomas autóctonos. Finalmente se evaluó el desempeño de métodos supervisados y no supervisados de clasificación y predicción respecto del diagnóstico

Palabras Clave: Microbioma-Cáncer-Secuenciación-Explotación de Datos-Evaluación Médica

1. Introducción

Los métodos de nueva generación para secuenciación de ADN posibilitan el análisis masivo y a bajo costo de las comunidades de microorganismos alojados en el intestino humano. El creciente interés médico que suscitan estos estudios se basa en la probada asociación de estados de riqueza y diversidad del microbioma con patologías importantes como el cáncer colorrectal sobre el cual se focaliza este artículo. En

trabajos anteriores [1] se ha dejado establecida una línea de procedimientos a efectuar sobre las lecturas desde que salen del secuenciador hasta que resultan procesadas en términos de explotación de datos. También se ha probado la potencialidad de estos métodos para caracterizar el microbioma [2]. Sin embargo, la elección de parámetros y algoritmos debe estar guiada por el criterio médico para el cual resulte útil la información aportada en términos clínicos de diagnóstico y evaluación. Este artículo se propone exhibir aspectos de la vinculación bioinformático- médica y ajustar la metodología hasta aquí desarrollada a efecto de hacer evidentes los aspectos clínicos de interés. Por primera vez se realiza el estudio sobre pacientes autóctonos, para los cuales la composición del microbioma varía de acuerdo a factores como tipo de alimentación, edad y localización geográfica. La tarea se realizó en el marco de un convenio firmado entre la Universidad Nacional de La Matanza y el Hospital Italiano de Buenos Aires, Sector de Coloproctología. A través del mismo se cuenta además con la inserción en la Large Bowel Microbiome Disease Network de la Universidad de Leeds, Inglaterra, lo que permite validar los procedimientos que se lleven a cabo.

Las tecnologías de nueva generación para la secuenciación de ADN han 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, que se ha considerado como un órgano adicional en las personas [2], varía 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 y que se asocia con enfermedades tales como la inflamación intestinal, el asma o los desórdenes mentales. En particular la disbiosis puede estar implicada en la carcinogénesis al ser iniciadora de procesos inflamatorios y su presencia da señal de inmunodepresión [3].

Algunos argumentos indirectos sugieren este rol potencial de la microbiota intestinal en la carcinogénesis colorrectal. El cáncer colorrectal 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 aparecen asociados el cáncer colorrectal y las variaciones de las frecuencias con que algunas especies bacterianas se encuentran en el microbioma [4] y [5]. A su vez la disminución en la diversidad total se ha vinculado con distintas patologías que incluyen el cáncer colorrectal, la obesidad, enfermedades autoinmunes y neurológicas [6]. 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 todas las secuencias obtenidas del microbioma (WGS), pueden establecer las vías metabólicas que potencialmente sigan los procesos celulares en el paciente [7]. Esto ha motivado un profundo interés en la comunidad médica que ha buscado avanzar en la comprensión, y eventualmente en el diagnóstico y pronóstico de enfermedades, utilizando estos métodos de análisis.

Al respecto hay que señalar no solo la tecnología de secuenciación sino también los desarrollos de algoritmos 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, y de aprendizaje sobre conjuntos de entrenamiento y testeo para clasificar microbiomas según los mismos principios.

En el proyecto Aplicación de Técnicas de Data Mining para Análisis del Microbioma Humano según Funcionalidades Metabólicas, desarrollado por el grupo en el período 2017-2018, se ha podido establecer una "pipeline", con varios pasos automatizados, para tratar las secuencias de ADN microbiómico. Comprende el tratamiento de las lecturas desde que salen del secuenciador hasta que resultan datos para explotación por técnicas estadísticas multivariadas y de aprendizaje supervisado y no supervisado, de forma de ponerlos al servicio de la interpretación médica. Estos procesos comienzan con el filtrado de las lecturas para quitar posibles contaminaciones con los reactivos utilizados en la secuenciación, continúan con el ensamblado en contigs, luego con el filtrado de las secuencias humanas que pudieran haber sido obtenidas también en la muestra y finalmente con la anotación taxonómica y funcional. Luego de esto la información debe disponerse de manera adecuada para iniciar el proceso de explotación de los datos que consiste en la aplicación de variadas técnicas estadísticas y de aprendizaje automático a efecto de establecer las características y patrones de comportamiento que puedan asociarse a la condición clínica de los pacientes. Para este trabajo se logró contar con muestras de materia fecal de pacientes autóctonos para iniciar así un estudio sobre las características locales de la enfermedad que se supone presentarán variaciones ligadas a dieta, condiciones de hábitat, etc. [8]

2. Materiales y Métodos

2.1 Muestras

Diseño:
Corte transversal.
1. 20 pacientes (10 con CCR y 10 controles) tratados por la Sección de Coloproctología del Hospital Italiano de Buenos Aires.
2. 15 pacientes (7 con CCR y 8 controles) tratados por la Sección de Coloproctología del Hospital Italiano de Buenos Aires.

Criterio de inclusión:
Casos: - Edad mayor a 18 años. - Adenocarcinoma de colon confirmado con histología.
Controles: - Edad mayor a 18 años - Ausencia de neoplasia colónica (adenocarcinoma y adenoma) confirmada por video colonoscopía completa, con Boston mayor a 6 (al menos 2 puntos por sector).
Criterio de exclusión: - Consumo de antibióticos o probióticos en los últimos 6 meses. - CCR en tratamiento - Antecedentes de cirugía colorrectal, CCR, radioterapia pélvica o quimioterapia. - Antecedentes familiares compatibles con síndromes de CCR hereditario - Enfermedad inflamatoria intestinal o enfermedad intestinal infecciosa. - Incapacidad de dar consentimiento informado.
Muestras empleadas en el estudio:
Muestra 1: materia fecal de 10 pacientes con CCR no tratado, material fecal de 10 voluntarios sanos que se sometieron a una colonoscopia por alguna razón y se haya demostrado que tienen un intestino normal en la colonoscopía.
Muestra 2: materia fecal de 7 pacientes con CCR no tratado, material fecal de 8 voluntarios sanos que se sometieron a una colonoscopia por alguna razón y se haya demostrado que tienen un intestino normal en la colonoscopía.
Mezcla de muestras 1 y 2: Se identificaron 216 géneros comunes entre la Muestra 1 y la Muestra 2. Con ellos y conservando el diagnóstico clínico efectuado se integró la mezcla de muestras con el objetivo de lograr una mayor representatividad y homogeneidad.

2.2 Secuenciación

Muestra 1: Se realizó con secuenciador Illumina HiSeq sobre la región V4 del gen 16S rRNA. Cada secuencia representa 150 pares de bases
Muestra 2: Se realizó con secuenciador Illumina MiSeq sobre las regiones V3 y V4 del gen 16S rRNA. Cada secuencia representa 300 pares de bases.

2.3 Procesamiento inicial

Ambas muestras fueron tratadas en una cadena de procesos establecida en trabajos anteriores [1]. La metodología empleada en estos procesos iniciales fue validada aquí, por comparación con trabajos similares realizados por grupos dentro de la Large Bowel Microbiome Disease Network, la cual integra el Hospital Italiano de Buenos Aires. Se importaron las lecturas del microbioma de cada paciente al software QIIME2 [9]. Luego se eliminó el ruido. Se filtraron las secuencias y se eliminaron las lecturas ambiguas o de baja calidad. A continuación, las distintas secuencias fueron alineadas contra los alineamientos de referencia para el gen 16S rRNA. Para cada metagenoma intestinal, se generó una tabla de frecuencias de las secuencias agrupadas en Unidades Taxonómicas Operacionales (OTU) y se confeccionó el árbol filogenético. En la Muestra 1, cuyas secuencias comprendieron solo la región V4 del gen, éstas se agruparon en 239 OTUs distintas, correspondientes al nivel taxonómico género. En la Muestra 2, más rica por contener las regiones V3 y V4 del gen, se pudieron identificar 370 taxones género.

2.4 Clustering

Se realizaron distintos experimentos de agrupamiento de pacientes a efecto de evaluar las posibilidades de la técnica en la clasificación clínica adecuada de los pacientes de acuerdo a su perfil microbiómico. Se realizaron pruebas de clustering jerárquico, con distancia euclídea, otras con agrupamiento no jerárquico por medio del algoritmo k-means, con distancia euclídea y encadenamiento promedio, variando el número inicial de centroides. Y finalmente se construyó "ad hoc" una distancia entre microbiomas que tiene en cuenta el peso de la diferencia de cada taxón entre pacientes sanos y enfermos [1]. En estos procesos se utilizó software INFOSTAT [10], WEKA [11] y desarrollos propios en lenguaje C para operar entre paquetes cambiando formatos y armar la matriz de distancias pesadas. Los agrupamientos fueron evaluados por el índice Silhouette.

2.5 Árboles de decisión

En relación con los métodos de aprendizaje automático, en base a los antecedentes de desempeño [12], se decidió entrenar y testear dos algoritmos de árboles de decisión. Por un lado, el C4.5 [13] disponible en Weka bajo el nombre J48 y por otro, el ensamble Random Forest, también incorporado a WEKA. Desde el punto de vista computacional se utilizaron matrices de confusión y curvas ROC para evaluar tanto el entrenamiento, realizado a partir de la Muestra 1, como el testeo, efectuado sobre la Muestra 2. La consideración comparativa de ambas muestras requirió la identificación de los taxones presentes simultáneamente en ambas. Se identificaron 216 géneros comunes con los cuales se trabajó en los dos tipos de árboles. Además, los mismos algoritmos se probaron con la mezcla de muestras 1 y 2. Así se seleccionó convenientemente un conjunto de entrenamiento de 18 pacientes y otro de testeo de 17. En todos los casos, se estableció como criterio relevante en términos clínicos que la clasificación fuera muy eficiente en la detección de pacientes enfermos y menos importante en cuanto a la verificación de los sanos.

3. Resultados obtenidos

Los primeros resultados obtenidos corresponden a los procesos iniciales realizados con QIIME2. Por ejemplo, la distribución estadística de frecuencias de OTUs o taxones se dispuso como exhibe la Tabla 1.

Tabla 1. Datos de la muestra 1, cantidad presente de cada taxón u OTU por paciente.

Taxón	D_0__Archae	D_0__Archae	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0__Bacter	D_0_
GCRFNG_AR-CC-A1	0	0	5	0	42	0	0	0	0	0	321	0	0	
GCRFNG_AR-CC-B2	0	0	0	0	22	0	0	0	0	0	73	0	4	
GCRFNG_AR-CC-C3	160	0	2	0	213	0	0	0	0	23	103	0	1	
GCRFNG_AR-CC-D4	0	0	0	0	359	1	0	0	0	0	77	0	0	
GCRFNG_AR-CC-E5	0	0	0	0	143	0	0	0	0	0	0	4	0	
GCRFNG_AR-CC-F6	0	0	10	0	1705	0	0	0	0	0	34	0	0	
GCRFNG_AR-CC-G7	58	0	0	0	3090	0	0	0	0	0	571	0	0	
GCRFNG_AR-CC-H8	0	0	0	0	89	0	0	0	0	0	86	0	0	
GCRFNG_AR-CC-I9	0	0	0	0	40	0	0	0	0	0	284	0	0	
GCRFNG_AR-CC-J10	49	64	141	0	64	0	4	0	0	0	134	0	0	
GCRFNG_AR-HV-A1	1	0	0	0	21	0	0	0	0	0	295	0	0	
GCRFNG_AR-HV-B2	10	0	0	0	95	0	0	0	12	0	55	0	0	
GCRFNG_AR-HV-C3	0	0	7	0	192	0	0	0	0	24	180	0	0	
GCRFNG_AR-HV-D4	0	0	0	0	180	0	0	11	0	0	114	0	10	
GCRFNG_AR-HV-E5	222	98	0	0	112	0	0	23	23	38	130	0	8	
GCRFNG_AR-HV-F6	0	0	16	0	128	0	0	0	0	5	227	0	0	
GCRFNG_AR-HV-G7	0	0	0	0	1105	0	0	0	0	4	95	0	9	
GCRFNG_AR-HV-H8	0	0	0	0	297	0	0	0	0	61	106	0	0	
GCRFNG_AR-HV-I9	0	0	0	0	3711	0	0	0	0	0	537	0	0	
GCRFNG_AR-HV-J10	1	0	13	1	48	0	8	0	0	0	66	0	0	

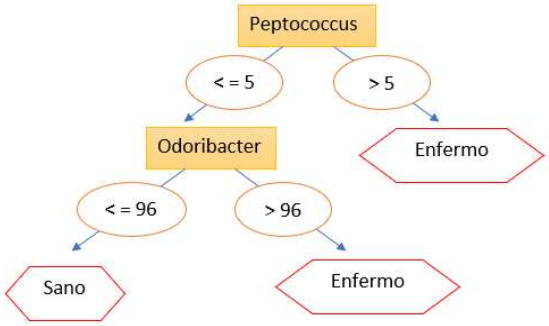
En la primera columna de cada tabla se anotan los pacientes detallados con un código, y la primera fila nombra cada uno de los taxones u Otus identificados. En la muestra 1, a nivel género, se pudieron identificar 239 Otus, y en la muestra 2, 368. Las últimas tres columnas corresponden a la clasificación, la edad y el sexo.

El clustering realizado con la muestra 1 arrojó resultados dispares. El método jerárquico fue poco adecuado para producir agrupamientos que se correlacionaran con la clasificación clínica por sano o enfermo. En cambio, K-means con distancia euclídea y encadenamiento promedio arrojó mejores resultados, aunque insuficientes para asegurar una clasificación adecuadamente correlacionada con el diagnóstico conocido. Esto se logró al establecer una distancia pesada "ad hoc". Para ello se consideró la media de las frecuencias de cada taxón teniendo en cuenta el diagnóstico clínico. Para cada taxón se calculó el valor absoluto de la diferencia entre la media de los pacientes sanos y la media de los enfermos. A cada resultado se lo dividió por la suma total de las diferencias y se lo multiplicó por un factor positivo constante para constituir el peso de cada taxón. Estos pesos se incorporaron al cálculo de la distancia entre pacientes de acuerdo a: $d = \sqrt{\sum_{j=1}^{239}(f_{ij} - f_{kj})^2 P_j}$. Con la matriz de las nuevas distancias se aplicó k-means para obtener ahora dos clusters. Se observó que los casos enfermos fueron todos bien clasificados, mientras que solo resultaron bien clasificados la mitad de los pacientes sanos. El test chi-cuadrado para evaluar la asociación entre la clasificación clínica y los clusters obtenidos arrojó un valor p <0.009 lo que indica que puede rechazarse la independencia entre ambas variables cualitativas. Los agrupamientos óptimos alcanzaron índices silueta de 0.49 para el cluster 1 que agrupó todos los casos enfermos y la mitad de los sanos y de 0.12 para el cluster 2. El índice silueta general fue de 0.4 lo que se consideró aceptable habida cuenta de la óptima clasificación de los casos enfermos.

Al realizar sobre la muestra 2 el agrupamiento por medio de k-means, con la distancia pesada y encadenamiento promedio se obtuvo un resultado parecido. El valor p para la prueba chi-cuadrado resultó p < 0.07 por lo cual puede rechazarse la hipótesis de independencia entre diagnóstico y cluster con ese nivel de significancia. El cluster 1 agrupó el total de los 7 casos enfermos y 5 de los voluntarios sanos, mientras que el cluster 2 se integró con 3 de los 8 casos sanos (37.5 %). El índice silueta del

agrupamiento total resultó de 0.73 y como ocurrió para la otra muestra fue mejor el índice silueta del cluster 1, 0.84 que el del cluster 2, 0.27.

El algoritmo J48 se corrió sobre la muestra 1 dividida en conjuntos de entrenamiento y testeo. Como el desempeño fue pobre, en este caso además se realizó una selección de atributos por medio de un procedimiento que establece un ranking de variables según la información que aportan a la variable de clasificación [14]. Así se seleccionaron solo 20 géneros para entrenar y testear. De ellos lo que mejor rankearon fueron el 119, Peptococcus , y 22, Odoribacter. Ambos le bastaron al modelo predictivo J48 para establecer, podando los otros, la regla de inferencia de la Figura 1.



Peptococcus <= 5
| Odoribacter <= 96: HV (10.0)
| Odoribacter > 96: CC (3.0)
Peptococcus >5: CC (7.0)

Figura 1. Diagrama de árbol determinado por las reglas de inferencia J48.

Solo el 30% de los casos fue bien clasificado en el testeo, lo que motivo el descarte del algoritmo en este trabajo.

A continuación, sobre las muestras homogeneizadas en los 216 taxones comunes se aplicó el ensamble Random Forest [15]. Se realizaron distintas experiencias. Se tomó como conjunto de entrenamiento, la muestra 1 de 20 pacientes, y se testeó con la muestra 2 de 15 pacientes. El porcentaje de casos de testeo bien clasificados fue del 60% pero lo importante es que el algoritmo detectó bien todos los casos enfermos, aunque solo clasificó adecuadamente a la cuarta parte de los sanos. El área bajo la curva ROC de testeo fue de 0.946 por lo que la diferencia con la de entrenamiento, que había clasificado bien todos los casos, es de 0.054 lo que revela un entrenamiento adecuado.

Se corrió también el algoritmo Random Forest sobre la mezcla de las muestras 1 y 2. En este caso se realizó una selección previa de atributos basada en el criterio de pesos ya utilizado en el clustering para calcular las distancias. Con 9 atributos para entrenar el ensamble el 64 % de los casos resultaron bien clasificados, pero aquí solo el 75 % de los enfermos fue clasificado como tal. La diferencia entre el área bajo las curvas ROC fue de 0.354 lo que revela el sobreentrenamiento a pesar de la poda de atributos efectuada.

Un resumen de los métodos aplicados y sus resultados se muestra en la Tabla 2.

Tabla 2. Desempeño de Algoritmos

Método	Algoritmo	M	Selección	%	%E/CC	%S/HV	Sil/DifARoc
Cluster:	Jerárquco	1	Dist Eucl	5	100	10	----------
Cluster	Kmeans	1	Dist Eucl	10	90	10	0.26
Cluster	Kmeans	1	Dist Pes	75	100	50	0.40
Cluster	Kmeans	2	Dist Pes	67	100	37	0.73
Ar. Dec	J48	1	Infogain	30	20	40	0,700
Ar. Dec	R. Forest	1y 2	Sin Selec	60	100	25	0.054
Ar. Dec	R. Forest	Me12	Pesos	64	75	56	0.354

El porcentaje de enfermos bien clasificados como tales se resume en la Tabla 2 como %E/CC y el porcentaje de pacientes sanos correctamente clasificados se simboliza por %S/HV

4. Conclusiones

Se ha logrado realizar toda la cadena de análisis necesaria para la determinación microbiómica por genes marcadores con pacientes autóctonos de la zona del AMBA. Se ha realizado la secuenciación de muestras de ADN de materia fecal, se han completado los procesos de filtrado, alineamiento y reconocimiento taxonómico siguiendo el método validado a nivel internacional. Durante la ejecución de esos procesos se han concretado también todos los enlaces necesarios relativos a cambios de formatos y presentaciones de la información lo cual, detallado parcialmente en trabajos anteriores [1], está aquí implícito. Así la información obtenida ha estado disponible para realizar pruebas de desempeño de algoritmos de explotación de datos en la determinación clínica. Respecto al clustering, se han dado resultados prometedores con la distancia pesada definida. Lo mismo ha ocurrido con la aplicación del ensamble de árboles de decisión Random Forest teniendo en cuenta la alta proporción de clasificación correcta de los pacientes enfermos. Resulta claro que deben realizarse ensayos más amplios utilizando muestras de mayor tamaño para afinar y confirmar la efectividad al utilizar estas técnicas para apoyar el diagnóstico. Sin embargo, tanto los clusters hallados con distancia pesada, como los ensayos con el ensamble de árboles han cumplido con el criterio general de mínimo error en la clasificación de los pacientes enfermos, lo que puede constituir una herramienta no invasiva para determinar la realización de otros estudios.

5. Referencias

1. Avila Laura, Santa María Victoria, López Luis, Soria Marcelo y Santa María Cristóbal.: Tratamiento de Secuencias de ADN y Clustering de Pacientes con Cáncer Colorrectal. WICC2020. El Calafate. (2020) https://wicc2020.unpa.edu.ar/
2. O'Hara AM, Shanahan F.: The gut flora as a forgotten organ. EMBO Rep. 2006 Jul;7(7):688–93. (2006)
3. Lopez, A et al.: Microbiota in digestive cancers: our new partner? Carcinogenesis, 1-10. doi:10.1093/carcin/bgx087 (2017)
4. Kosumi K, Hamada T, Koh H, Borowsky J, Bullman S, Twombly TS, et al.: The Amount of Bifidobacterium Genus in Colorectal Carcinoma Tissue in Relation to Tumor Characteristics and Clinical Outcome. Am J Pathol [Internet]. 2018 Sep 20; (2018) Available from: http://dx.doi.org/10.1016/j.ajpath.2018.08.015
5. Youssef O, Lahti L, Kokkola A, Karla T, Tikkanen M, Ehsan H, et al.: Stool Microbiota Composition Differs in Patients with Stomach, Colon, and Rectal Neoplasms. Dig Dis Sci [Internet]. 2018 Jul 11; Available from: http://dx.doi.org/10.1007/s10620- 018-5190-5
6. Shreiner AB, Kao JY, Young VB.: The gut microbiome in health and in disease. Curr Opin Gastroenterol; 31(1):69–75. (2015)
7. 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, Article number: 4139. (2018) www.nature.com/scientificreports
8. Taylor M, Wood HM, Halloran SP, Quirke P.: Examining the potential use and longterm stability of guaiac faecal occult blood test cards for microbial DNA 16S rRNA sequencing. J Clin Pathol. 2017 Jul;70(7):600–6. (2017)
9. Bolyen E, et al.: Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nature Biotechnology 37: 852–857. (2019) https://doi.org/10.1038/s41587-019-0209-9
10. Di Rienzo J.A., Casanoves F., Balzarini M.G., Gonzalez L., Tablada M., Robledo C.W.: InfoStat versión 2018. Grupo InfoStat, FCA, Universidad Nacional de Córdoba, Argentina. (2018) URL http://www.infostat.com.ar
11. https://www.cs.waikato.ac.nz/ml/weka/Witten_et_al_2016_appendix.pdf
12. 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 1:11 (2013)
13. Quinlan, J.R.: C4.5 Programs for Machine Learning, San Mateo, CA: Morgan Kaufmann (1992)
14. Eibe Frank, Mark A. Hall e Ian H. Witten.: El banco de trabajo WEKA. Apéndice en línea para "Minería de datos: herramientas y técnicas prácticas de aprendizaje automático", Morgan Kaufmann, cuarta edición. (2016)
15. Breiman, Leo.: Random Forests. Machine Learning 45 : 5–32. doi:10.1023/A:1010933404324. (2001)

Congreso Argentino de Ciencias de la Computación

San Justo, Octubre de 2020

Se certifica que

Laura Ávila, Victoria Santa María, Luis López, Cristóbal Raúl Santa María y Marcelo Soria

han participado como Autores del artículo 13384 *"Análisis del Desempeño de Clustering y Árboles de Decisión en la Evaluación Clínica de Microbiomas de Pacientes con Cáncer Colorrectal"*, aceptado en el XXVI Congreso Argentino de Ciencias de la Computación, organizado por la Universidad Nacional de La Matanza, del 5 al 9 de octubre de 2020.

Lic. Patricia Pesado
Coordinadora Titular de RedUNCI

Mg. Jorge Eterovic
Decano DIIT UNLaM

DIIT

Universidad Nacional de La Matanza

Minería de Datos del Microbioma en Pacientes con Cáncer Colorectal

Cristóbal Santa María UNLaM csantamaria@unla.edu.ar
*Laura Ávila UNLaM*Laura_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.

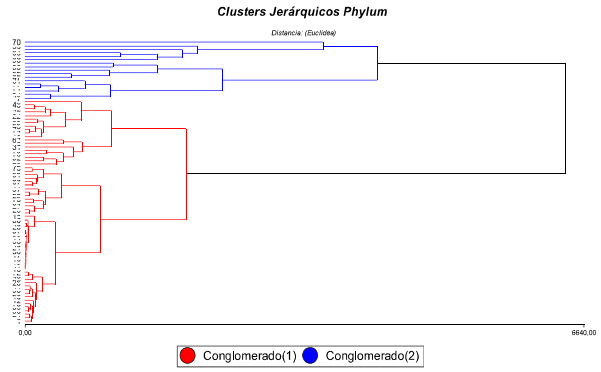


Figura 1. Clusters Jerárquicos Phylum

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

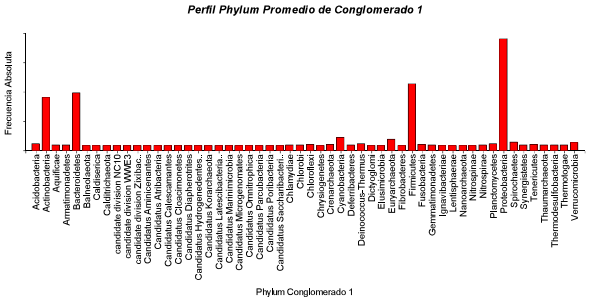


Figura 2. Perfil Phylum de Promedios-Conglomerado 1

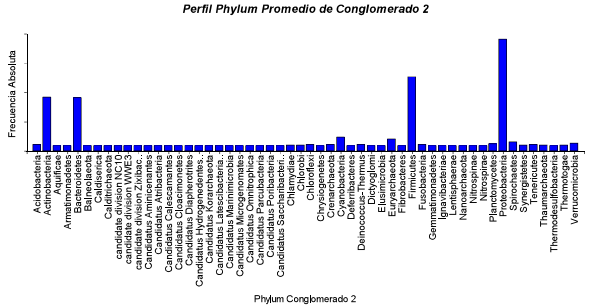


Figura 3. Perfil Phylum de Promedios-Conglomerado 2

- 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.

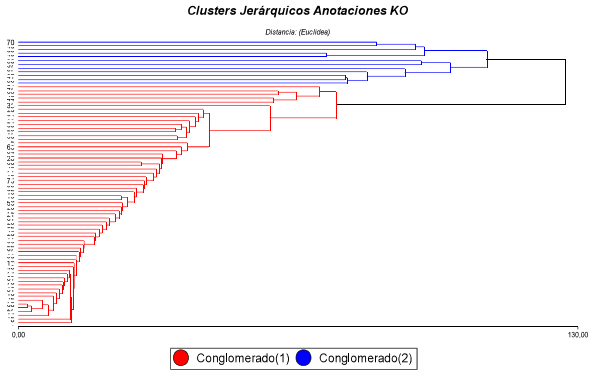


Figura 4. Clusters Jerárquicos Anotaciones KO

- 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.

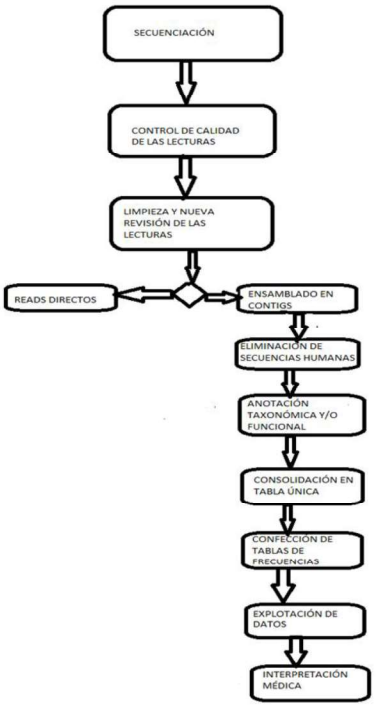


Figura 5. Flujo de Datos

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

CONAIISI

VII Congreso Nacional de Ingeniería Informática - Sistemas de Información

2019

San Justo, 5 de diciembre de 2019

Se certifica que **Cristóbal Raúl Santa María, Laura Avila, Victoria Santa María, Luis López, Marcelo Abel Soria** han participado como Autores del artículo 72 *"Minería de datos del microbioma en pacientes con cáncer colorectal"*, aceptado en el VII Congreso Nacional de Ingeniería Informática – Sistemas de Información, CONAIISI 2019, realizado los días 14 y 15 de noviembre en la Universidad Nacional de La Matanza.

Ing. Claudio D'Amico
Coord. Gral. CONAIISI

Dr. Carlos Neil
Coordinador RIISIC

Mg. Jorge Eterovic
Decano DIIT

UNLaM

confedi
Consejo Federal de Decanos de Ingeniería
República Argentina

DIIT
Departamento de Ingeniería e Investigaciones Tecnológicas

RIISIC

Universidad Nacional de La Matanza

DIIT
Departamento de Ingeniería e Investigaciones Tecnológicas

Mejora de las Estrategias Pedagógicas

San Justo, 15 de diciembre de 2020

Se certifica que

Cristóbal Santa María

DNI: 11.360.708

participó como Expositor en el *"IV Encuentro del Programa MEP -Mejora de las Estrategias Pedagógicas-"* (Resolución de Rectorado Nº 294), dictado por la Dra. Bettina Donadello, el 4 de diciembre del corriente, en esta Casa de Altos Estudios.

Dra. Bettina Donadello
Secretaria de Investigaciones

Mg. Ing. Jorge Eterovic
Decano

UNLaM
Universidad
Nacional de
La Matanza

Instituto Universitario
Hospital Italiano

CONVENIO MARCO DE COOPERACION
ENTRE
INSTITUTO UNIVERSITARIO ESCUELA DE MEDICINA DEL HOSPITAL ITALIANO Y LA UNIVERSIDAD NACIONAL DE LA MATANZA

Entre la **Instituto Universitario Escuela de Medicina del Hospital Italiano,** en adelante **"IUHI"** representada en este acto por su Rector, Prof. Dr. Marcelo Fernando Figari con D.N.I. N° 13.530.740 y domicilio legal en Potosí 4240, CABA, y la **UNIVERSIDAD NACIONAL DE LA MATANZA**, en adelante **"la Universidad"**, representada en este acto por el Sr. Rector Prof. Dr. Daniel Eduardo Martínez con D.N.I. N° 10.633.043 y domicilio legal en Florencio Varela 1903, San Justo, Provincia de Buenos Aires, acuerdan celebrar el presente convenio:

Ambas partes acuerdan:

- Que la mutua complementación y cooperación sirven a su respectivo desarrollo institucional incrementando sus capacidades de investigación tecnológica y de difusión y preservación de la cultura.

- Que este modo de vinculación permitirá un mejor servicio a las necesidades de la Comunidad, reconociendo las mismas como así también las cuestiones relativas a la solidaridad social, al empleo y la producción.

CLAUSULA PRIMERA

1.1 Ambas partes convienen en establecer relaciones de complementación, cooperación y asistencia recíproca de carácter académico, cultural, tecnológica y de servicio.

1.2 Dichas relaciones se efectivizarán por la adopción de medidas de coordinación y acción en común en todas las Áreas propias de su incumbencia, toda vez que las circunstancias lo aconsejen y permitan.

1.3 Las Instituciones signatarias manifiestan su voluntad de llevar a cabo, entre otras, las siguientes acciones:

1.3.1 Actuar cada una como "Organismo Asesor" de la otra en relevamiento y resolución de problemas sobre temas de su competencia.

1.3.2 Colaborar en proyectos de investigación y desarrollo que la contraparte tenga en ejecución, intercambiando información y personal idóneo.

1.3.3 Organizar conferencias, seminarios y cursos relativos a temas de interés de alguna de las partes.

1.3.4 Desarrollar programas de formación profesional y atención comunitaria que permitan satisfacer las demandas generadas por la comunidad, comprometiendo la participación de los actores sociales en su gestión.

CLAUSULA SEGUNDA

Ambas partes se comprometen a:

2.1 Reconocer como funciones normales de su personal docente y/o técnico el cumplimiento de las tareas que se les asigne en virtud del presente Convenio, sin que ello implique obligación pecuniaria alguna de la celebrante, salvo acuerdo expreso en contrario.

2.2 Prestar facilidades de acceso a los servicios académicos, científicos, tecnológicos y culturales a los docentes, graduados, estudiantes y personal técnico o administrativo de la cocelebrante.

CLAUSULA TERCERA

Ambas Instituciones concuerdan en abrir los campos de intercambio a todas las disciplinas o especialidades propias de cada una de ellas.

CLAUSULA CUARTA

El presente Convenio no debe interpretarse en el sentido de haber generado una relación legal o financiera entre las partes. Las condiciones particulares relativas al financiamiento, organización y ejecución de las actividades a desarrollar como así también las cuestiones de índole judicial respecto de desacuerdos o diferencias que pudieran originarse serán establecidas; para cada caso en particular, en Protocolos adicionales al presente.

CLAUSULA QUINTA

Cualquiera de las partes podrá denunciar el presente Convenio mediante comunicación escrita dirigida a la otra, con 6 (seis) meses de anticipación, sin que la denuncia afecte acciones pendientes o en curso de ejecución que no fuesen expresamente rescindidas por ambas Instituciones.

UNLaM
Universidad
Nacional de
La Matanza

Instituto Universitario
Hospital Italiano

01

CLAUSULA SEXTA

El presente Convenio tendrá vigencia por un período de 4 (cuatro) años a contar desde la fecha de suscripción, pudiendo renovárselo por iguales períodos sucesivos mediante el intercambio de notas ratificatorias oficializadas con una antelación de 60 (sesenta) días corridos previos a la expiración de cada período.

En muestra de conformidad se firman 2 (dos) ejemplares del presente, de idéntico tenor y a un solo efecto en San Justo, a los 20.. días del mes de MARZO de 2019

Por el Instituto Universitario Escuela de Medicina del Hospital Italiano

Dr. Marcelo F. Figari
Rector

Prof. Marcelo Figari
Rector
Instituto Universitario
Escuela de Medicina del Hospital Italiano

Por la Universidad Nacional de la Matanza

Dr. Daniel E. Martínez
Rector

Instituto Universitario

Escuela de Medicina del Hospital Italiano

Departamento de Posgrado

HOSPITAL ITALIANO
de Buenos Aires

Por cuanto

Santa María, Cristóbal Raúl

Ha asistido a la Actividad:

II Jornadas del Instituto de Medicina Traslacional e Ingeniería Biomédica (IMTIB) y I Jornada de Microbioma: Aspectos Prácticos de la Investigación Traslacional

Llevada a cabo el día ***03 de octubre de 2019***
con un total de ***5*** *horas*

Se extiende el presente certificado de asistencia, en Buenos Aires, Argentina

Dr. Carlos Alberto Vaccaro
Director de la jornada

Dr. Marcelo Risk
Director de la jornada

Dr. Adrián Gadano
Director de la jornada

Código	FPI-009
Objeto	Guía de elaboración de Informe final de proyecto
Usuario	Director de proyecto de investigación
Autor	Secretaría de Ciencia y Tecnología de la UNLaM
Versión	5
Vigencia	03/9/2019

Cristóbal Raúl Santa María

.

Lugar y fecha: 25 de febrero de 2021