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	

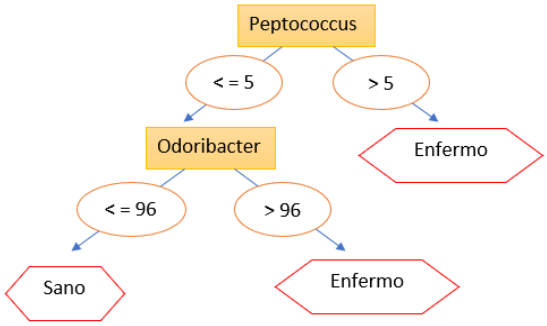
En la primera columna de cada tabla se anotan los pacientes detallados con un código, y la primera fila nombra cada uno de los taxones u Otus identificados. En la muestra 1, a nivel género, se pudieron identificar 239 Otus, y en la muestra 2, 368. Las últimas tres columnas corresponden a la clasificación,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