Síntesis informativa sobre actividades de investigación en el DIIT
Resúmenes Didácticos de los trabajos originales

Título del Proyecto: Explotación de Datos del Microbioma de Pacientes con Cáncer Colo-rectal

Código: C220
Director: Cristóbal Santa María **e-mail: csantamaria@unlam.edu.ar**
Co-Director: Luis López **e-mail: llopez@unlam.edu.ar**
Integrantes:
Investigadores: Laura Ávila, Victoria Santa María, Marcelo Soria, Ariel Cacho Mendoza y Pablo Martínez

Síntesis del contenido: 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

Temario de presentación: 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 se basa en la probada asociación de estados de riqueza y diversidad del microbioma con patologías importantes como el cáncer colorrectal. Por primera vez se realizó el estudio sobre pacientes autóctonos, en el marco de un convenio firmado entre la Universidad Nacional de La Matanza y el Hospital Italiano de Buenos Aires. Se contó además con la inserción en la Large Bowel Microbiome Disease Network de la Universidad de Leeds, Inglaterra, para validar los procedimientos. El microbioma humano no es otra cosa que la comunidad de microorganismos presentes en el cuerpo humano. La composición del microbioma varía según el estilo de vida, la dieta y su genotipo, pero es estable dentro de una misma persona. Las modificaciones de tipo permanente pueden implicar distintas patologías y en particular cáncer colorrectal. Así la exploración del microbioma puede colaborar en el diagnóstico y pronóstico de la enfermedad.

Metodología del trabajo desarrollado:
a- Muestra 1: materia fecal de 10 pacientes con CCR (Cáncer-Colo-Rectal) 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. 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.

b- 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.

c- Procesamiento inicial. Ambas muestras fueron tratadas en una cadena de procesos establecida en trabajos anteriores. Se importaron las lecturas del microbioma de cada paciente al software QIIME2. 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.

d- Clustering. Se realizaron distintos experimentos de agrupamiento de pacientes a efecto de la clasificación clínica de los pacientes. Además, 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

e- Árboles de decisión

Se decidió entrenar y testear dos algoritmos de árboles de decisión. Por un lado, el J48 y por otro, el ensamble Random Forest. Se utilizaron matrices de confusión y curvas ROC para evaluar tanto el entrenamiento y el testeo.

Desarrollo y resultados obtenidos:

El clustering mediante el método 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”. Con la matriz de las nuevas distancias se obtuvieron dos clusters. Se observó que los casos enfermos fueron todos bien clasificados. El test para evaluar la asociación entre la clasificación clínica y los clusters obtenidos indicaron que puede rechazarse la independencia entre ambas variables cualitativas. Los agrupamientos óptimos alcanzaron índices de buen desempeño. Al realizar sobre la segunda muestra el agrupamiento por medio de k-means, con la distancia pesada y encadenamiento promedio, se obtuvo un resultado parecido.

Se realizaron distintas experiencias con el ensamble Random Forest. 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 entrenamiento se juzgó 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. Se observó sobreentrenamiento a pesar de la poda de atributos efectuada.

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, 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.

Publicaciones y/o transferencias empleadas:

- Cristóbal Santa María, Laura Ávila, Victoria Santa María, Luis López y Marcelo Soria “Minería de datos del microbioma en pacientes con cáncer colo-rectal”. 2019. CONAIISI
- Ávila, Laura | Santa María, Victoria | López, Luis | Soria, Marcelo | Santa María, Cristóbal “Tratamiento de secuencias de ADN y clustering de pacientes con cáncer colo-rectal”. 2020. WICC
- Ávila, Laura | Santa María, Victoria | López, Luis | Santa María, Cristóbal/ Marcelo Soria. “Evaluación clínica de microbiomas de pacientes con cáncer colo-rectal” 2020. CACIC
- Cristóbal Santa María. “Clustering y árboles de decisión en pacientes con cáncer colorectal”. 2020. IV Encuentro del Programa MEP – UNLAM
- Agencia CTyS-Unlam “Analizan el vínculo entre el microbioma y el Cáncer de Colon” Entrevista. 2020

Bibliografía Utilizada

1. Lopez, A et al.: Microbiota in digestive cancers: our new partner? Carcinogenesis, 1-10. doi:10.1093/carcin/bgx087 (2017)
2. 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
3. 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