CLUSTERING Y ENSAMBLES DE ÁRBOLES DE DECISIÓN APLICADOS SOBRE EL MICROBIOMA HUMANO

Cristóbal R. Santa María. Departamento de Ingeniería. UNLAM
Marcelo Soria. Facultad de Agronomía Cátedra de Microbiología UBA
Victoria Santa María. Facultad de Medicina. Instituto A. Lanari. UBA
Fernando Galanternick. Facultad de Medicina. Instituto A. Lanari. UBA
Florencio Varela 1903 San Justo Pcia. de Buenos Aires
54-011-44808952
csantamaria@unlam.edu.ar
soria@agro.uba.ar
vctrsntmr@gmail.com
fgalanternik@gmail.com

RESUMEN

El objetivo general es desarrollar un algoritmo de clasificación de estadíos de desarrollo de cáncer de colon y enfermedad de Crohn basado en la información aportada por el ADN del microbioma humano. A partir de la secuención del ADN microbiano presente en el intestino, cada secuencia genética es una instancia en una base de datos sobre la que es posible aplicar procedimientos de aprendizaje no supervisado para agrupar las secuencias correspondientes a un gen marcador por especies u otros taxones más generales. Tales categorizaciones tienen un sesgo, respecto de la caracterización clínica, que es producto de la secuenciación misma y de las técnicas que se aplican previas al agrupamiento. Conviene entonces explorar agrupamientos de todas las secuencias genéticas, y no ya solo las de un gen marcador, de acuerdo a la función que les corresponda en el metabolismo y que resulta distinta en la salud y en cada estadío de la enfermedad. Se espera que el estudio realizado interrelacionando ambos tipos de agrupamientos proporcione categorías de clasificación estables y compatibles con las caracterizaciones clínicas de la enfermedad. Con tales categorías se intentará aplicar métodos de aprendizaje supervisado como ensambles de árboles de decisión para obtener un clasificador que colabore en la clínica de prevención, diagnosis o prognosis. Se pretende además elaborar una "pipeline" para investigar la aplicación de técnicas de data mining al microbioma humano en el caso de una enfermedad en general.

Palabras Clave: ADN, Microbioma, Cluster, Ensamble, Predicción, Pipeline

CONTEXTO

La línea de trabajo que aquí se presenta se inscribe en el proyecto de investigación de técnicas de minería de datos aplicadas sobre bases de secuencias de ADN correspondientes a la colección de microorganismos que pueblan el cuerpo humano. El proyecto tiene la finalidad de estudiar la interacción recíproca entre esa colección denominada microbioma humano y los estados de salud y enfermedad del portador humano. Se espera que el uso de estos procedimientos suministre valiosos elementos de apreciación clínica para prevenir, diagnosticar y pronosticar enfermedades

INTRODUCCIÓN

Hay dos formas posibles de encarar los procedimientos de data mining sobre bases de datos de ADN . Por un lado pueden utilizarse las secuencias correspondientes a un solo gen, presente en todos los individuos, que haya conservado su estructura principal a través de la evolución biológica. Esta es la llamada técnica del gen marcador, y si se mide la similitud entre secuencias por un modelo de distancia evolutiva

adecuado, pueden formarse distintos clusters que agrupen las secuencias por proximidad biológica. Como cada secuencia del gen marcador representa a un individuo distinto, si se elige para el clustering un umbral de distancia adecuado, los individuos de una misma especie irán a parar al mismo cluster. Con un umbral menos exigente los agrupamientos corresponderán a categorías más generales como género o familia. En el caso del microbioma humano se sabe que su riqueza y diversidad es decir la cantidad de clusters y el número de individuos que integra cada agrupamiento, varía según el estado de salud o el estadío de desarrollo de una enfermedad. De esta forma una variación en esos parámetros podría indicar la presencia de una patología. Sin embargo la riqueza y distribución del microbioma varía mucho de individuo a individuo, de edad a edad o entre razas. Además debido a las características del proceso químico de secuenciación del ADN y a los procesos previos al agrupamiento, puede ocurrir que la cantidad y distribución de abundancia de los clusters formados, denominados Unidades Taxonómicas Operacionales, no represente en forma adecuada las categorizaciones clínicas de una dada enfermedad. Es aquí donde se hace necesario recurrir a un procedimiento más general que consiste en considerar todas las secuencias genéticas obtenidas, y no sólo las del gen marcador, y agruparlas de acuerdo a la función que le corresponda a cada gen en el metabolismo. Estos agrupamientos resultan más estables de individuo a individuo que los que se obtienen utilizando el gen marcador. Como las vías metabólicas varían entre el estado de salud y los de la enfermedad, el número de clusters y la cantidad de genes por cluster de una muestra revelaría con más propiedad el estadío clínico de un paciente. A su vez se considera que el estudio realizado interrelacionando ambos tipos de agrupamientos proporcionará categorías de clasificación aún más estables y compatibles con las caracterizaciones clínicas de los estadíos de salud y enfermedad. Es decir que el trabajo se propone hallar las categorías, obtenidas computacionalmente por aprendizaje no supervisado, que mejor se adecuen a la clasificación que establece la clínica en la evolución del cáncer de colon y de la enfermedad de Crohn. Una vez logrado esto se intentará utilizar la categorización hallada para aplicar métodos de aprendizaje supervisado tales como ensambles de árboles de decisión para obtener un clasificador que colabore en la clínica de prevención, diagnosis o prognosis. A partir de un conjunto de datos genéticos de pacientes clasificados según las categorías obtenidas, se propone entrenar y testear un procedimiento que clasifique casos para colaborar así en la evaluación clínica de nuevos pacientes. Se pretende además elaborar como resultado de los pasos descriptos una "pipeline" para investigar la aplicación de técnicas de data mining al microbioma humano en el caso de otras enfermedades.

Microbioma y cáncer de colon

En Argentina el cáncer de colon tiene una alta prevalencia, constituye la segunda causa de muerte por cáncer y el 2.1% de las muertes totales. Gracias a las múltiples medidas de prevención y métodos de screening su mortalidad se encuentra en descenso no así su incidencia que aumenta progresivamente. La función primordial del colon es el reciclado de nutrientes y depende del microbioma colónico, la motilidad del colon y la absorción y secreción de la mucosa. La flora bacteriana es esencial para evitar la pérdida innecesaria de líquidos,

electrolitos, nitrógeno y energía. El microhábitat del colon compuesto por la luz, la capa de mucina y la superficie mucosa se encuentra colonizado por centenares de bacterias diferentes (Guarner y Malagelada, 2003).

Numerosos factores han sido considerados importantes en la generación del cáncer colorectal y ciertas condiciones clínicas son consideradas predecesoras de la enfermedad. En pacientes con la enfermedad de Crohn, por ejemplo, existe un riesgo aumentado hasta 20 veces con respecto a la población en general para el desarrollo de cáncer. La mayoría de los cánceres de colon se desarrollan a partir de un pólipo precursor o lesión premaligna (adenoma tubular - adenoma velloso - adenoma tubulovelloso – adenoma serrato). Se ha observado que las dietas ricas en grasas y pobres en fibras favorecen el desarrollo de cáncer de colon. También se asocia a la ingestión excesiva de alcohol, las deficiencias de calcio y el tabaquismo entre otras. Además, los cambios en la flora microbiana del intestino tienen un papel destacado en la tumorigenésis. En modelos murinos de cáncer de colon se observó un enriquecimiento de grupos bacterianos afiliados a los géneros Bacteroides, Odoribacter y Akkermansia, con respecto a animales sanos y un aumento de grupos pertenecientes a las familias Prevotellaceae y Porphyromonadaceae. La colonización de los intestinos de ratones libres de gérmenes con la microbiata de ratones con tumor indujo un aumento en la tumorigenésis en el colon comparado con animales colonizados con el microbioma intestinal de animales sanos (Zacular y cols. 2013). En humanos se observaron cambios en la riqueza relativa de varios grupos bacterianos, especialmente una reducción de bacterias productoras en pacientes con cáncer de colon (Weir y cols., 2013). El butirato constituye un sustrato fundamental para las células epiteliales del colon, y además presenta efectos tróficos específicos sobre los colonocitos normales, frenando el crecimiento de los colonocitos neoplásicos.

Microbioma y enfermedad de Crohn

La enfermedad de Crohn (EC) es una enfermedad inflamatoria intestinal crónica recidivante, caracterizada por el compromiso transmural del intestino, que puede afectar todo el tubo digestivo, desde la boca hasta el ano, y que afecta a dicho órgano en forma interrumpida (deja espacios de mucosa sana entre las lesiones). Las consecuencias de dicho proceso inflamatorio intestinal pueden ser estenóticas, por la fibrosis resultante, o bien microperforaciones, con formación de trayectos fistulosos entre distintos órganos intraabdominales. Esta enfermedad es un factor de riesgo de cáncer de colon. Los datos en nuestro país son limitados, pero según estadísticas de los Estados Unidos, la incidencia se encuentra entre 3,1 a 20,2 casos por 100.000 personas/año, y la prevalencia es de 201 cada 100.000 personas. Existe evidencia que sugiere que la incidencia y prevalencia a nivel mundial se encuentra en ascenso. (Modolecky y cols., 2012).

La evidencia actual de las enfermedades inflamatorias intestinales sugiere que se producen por una respuesta inflamatoria inapropiada hacia la microbiota intestinal en personas susceptibles genéticamente. Se han encontraron alteraciones de la composición dominante de la microbiota intestinal en pacientes con enfermedad de Crohn recientemente diagnosticada con sobrerrepresentación de Escherichia/Shigella y disminución de Faecalibacterium (Thorkildsen y cols., 2013). Por otra parte, en un análisis integrado y comparativo del metagenoma

y metaproteoma intestinal en pacientes con enfermedad de Crohn y controles sanos, se evidenciaron clasificaciones específicas vinculadas con los estados de enfermedad y numerosas vías metabólicas asociadas (Erickson y cols., 2013).

LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

La elección de técnicas de data mining para enfrentar los desafíos de la clasificación multicategorial que pueden presentar los estudios del microbioma requiere, en primer lugar, distinguir entre las dos clases típicas de métodos de aprendizaje automático. Por un lado los no supervisados que se emplean, por ejemplo, cuando se trata de medir la diversidad α que involucra la riqueza y abundancia de taxones presentes en una muestra. En tales casos se suele aplicar alguna variante de clustering jerárquico (Everitt y cols. 2001) y los coeficientes de riqueza resultante pueden utilizarse como categorías de clasificación en un estudio posterior de muestras y comunidades aplicando aprendizaje supervisado (Knights D y cols. 2011) La diversidad β permite medir la forma en que una determinada estructura comunitaria cambia de una comunidad a otra. Esta es una cuestión importante en el caso del cuerpo humano pues las comunidades microbianas alojadas en distintos órganos comúnmente no comparten las mismas especies aún con estructuras o funciones análogas. En tal caso el análisis de componentes principales (PCA) abre la posibilidad de utilizar algunas características comunes para formar clusters por órgano o por factores ambientales (Knights D y cols. 2011). El principal propósito del aprendizaje supervisado es estructurar un modelo desde un conjunto de datos ya clasificados que pueda predecir la clasificación correcta de datos no clasificados. Es el caso de la prevención del cáncer en donde la expresión de los perfiles genéticos puede prevenir o anticipar tipos y estadíos de la enfermedad. Estos algoritmos involucran árboles de decisión y ensambles de los mismos como Random Forest (Breiman L 2001). A través de las curvas ROC se logra optimizar el error esperado de predicción de tales procedimientos por comparación del desempeño sobre conjuntos de secuencias de entrenamiento y testeo. Se han realizado pruebas iniciales sobre el microbioma corporal completo, sobre el de la piel, brazos, manos y dedos en personas vivas y también con objetivos forenses (Knights D y cols. 2011). El trabajo más amplio en cuanto a comparación de métodos de clasificación ha sido publicado en 2013 (Statnikov y cols. 2013). Se utilizaron 8 conjuntos compuestos por un total de 1802 muestras de secuencias del gen 16s rRNA correspondientes a distintos subconjuntos del microbioma humano tales como la cavidad oral, el esófago, el canal auditivo, las heces, axilas, el estómago y otras en situaciones de salud o enfermedad como psoriasis, reflujo esofágico o adenocarcinoma esofágico etc. Las OTUs se armaron con un umbral de disimilaridad del 3%. y para evaluar el desempeño de los distintos algoritmos se usaron curvas ROC. Uno de los algoritmos que resultó más eficiente desde el punto de vista informático fue precisamente el ensamble Random Forest.

RESULTADOS Y OBJETIVOS

El trabajo ha comenzado relevando las fuentes de datos disponibles. En el repositorio SRA (Sequence Read Archive, http://www.ncbi.nlm.nih.gov/Traces/sra/) mantenido por el NCBI se han desarrollado búsquedas durante febrero de 2014. Sobre muestras relacionadas con enfermedad de Crohn se hallaron 341 registros provenientes de siete

relevamientos hechos en humanos, y para cáncer de colon se computaron más de 700 registros, algunos provenientes de pacientes humanos y otros de modelos experimentales en animales de laboratorio, que también resultan conveniente para las tareas propuestas. Para ambas enfermedades existen datos obtenidos a partir del gen 16S rRNA o secuencias obtenidas de genomas totales.

El estudio apunta a ajustar la aplicación de algoritmos no supervisados y supervisados en el caso concreto de las enfermedades analizadas evaluando no solo su desempeño computacional sino compatibilizando las categorías obtenidas con las definidas en forma clínica, poniendo a salvo la estabilidad de las mismas frente al sesgo introducido por el proceso de las secuencias. A su vez se intenta diseñar el modelo clasificador como herramienta real que colabore en la prevención, diagnóstico y pronóstico de dichas patologías. Así se tratará de estructurar un camino de trabajo luego aplicable a estudios y desarrollos sobre otras patologías.

BIBLIOGRAFÍA

-Breiman L. (2001) Random Forest. Machine Learning 45. 5-32

-Erickson AR, Cantarel BL, Lamendella R, Darzi Y, Mongodin EF, y cols. (2012) Integrated Metagenomics/Metaproteomics Reveals Human Host-Microbiota Signatures of Crohn's Disease. PLoS ONE 7(11): e49138. doi:10.1371/journal.pone.0049138

-Everitt, B, Landau, S, Leese, M. (2001) Cluster Analysis. Fourth Edition. Arnold.

-F. Guarner, J.R. Malagelada, (2003) Gut flora in health and disease. The Lancet, 361(9356):512-519. http://dx.doi.org/10.1016/S0140-6736(03)12489-0.

-Knights D. Costello E K. y Knight R. (2011) Supervised classification of human microbiota. FEMS Microbiol Rev 35 343-359

-Molodecky N.A., Soon I.S., Rabi D.M., y cols. (2012) Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology 2012(142):46-54.

-Statnikov A. Henaff M. Narendra V. Konganti K. Li Z. Yang L. Pei Z. Blaser M. Aliferis C y Alekseyenko A. (2013) A comprehensive evaluation of multicategory classification methods for microbiomic data. Microbiome 2013 1:11

-Thorkildsen LT, Nwosu FC, Avershina E, Ricanek P, Perminow G, Brackmann S, Vatn MH, Rudi K. (2013) Dominant fecal microbiota in newly diagnosed untreated inflammatory bowel disease patients. Gastroenterol Research and Practice. 2013:636785. doi: 10.1155/2013/636785.

-Weir TL, Manter DK, Sheflin AM, Barnett BA, Heuberger AL y cols (2013). Stool Microbiome and Metabolome Differences between Colorectal Cancer Patients and Healthy Adults. PLoS ONE 8(8): e70803. doi:10.1371/journal.pone.0070803

-Zackular, JP. Baxter NT. Iverson KD. Sadler WD. Petrosino JF. Chen GY. Y Schloss PD. (2013) The Gut Microbiome Modulates Colon Tumorgenesis. mBio 4(6):e00692-13