Técnicas de Minería de Datos Aplicadas al Procesamiento de ADN de Comunidades Microbiológicas

Cristóbal R. Santa María*, Victoria Santa María**, Fernando Galanternick**, Luis López*
Juan Otaegui*, Marcelo Soria***
*DIIT-UNLaM, **Instituto Lanari-FMed-UBA, ***FAUBA
Florencio Varela 1903 San Justo Pcia. de Buenos Aires 54-011-44808952
csanta_maria@ing.unlam.edu.ar
vctrsntmr@hotmail.com
fgalanternik@gmail.com
llopez@ing.unlam.edu.ar
soria@agro.uba.ar
juancarlosotaegui@yahoo.com.ar

Resumen

Se expone la línea de investigación que lleva adelante el Grupo de Investigación y Desarrollo en Data Mining del Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLaM. Se detallan los resultados del proyecto de investigación "Data Mining y Simulación en Evaluaciones de Biodiversidad", C141 del Programa de Incentivos, y las perspectivas de un nuevo proyecto, "Aplicaciones de Data Mining al estudio del Microbioma Humano", que se inicia dentro del mismo programa institucional.

Las modernas técnicas de secuenciación de ADN transforman su estructura química en secuencias informáticas de símbolos cada una de las cuales puede ser vista como una instancia de una base de datos. Es posible entonces aplicar métodos para clasificar casos y predecir patrones de comportamiento de forma similar a como se lo hace sobre otros dominios. Dentro de esta línea de trabajo se desarrolló un algoritmo que permite evaluar la cantidad de especies distintas en una comunidad microbiana, mejorando la eficiencia de otras estimaciones estadísticas a partir de muestras. Actualmente se trabaja en las formas de agrupamientos (clustering) que resulten compatibles con la evaluación clínica del metagenoma humano (microbioma), el cual sufre importantes variaciones en presencia de patologías. Se pretende desarrollar un clasificador de enterotipos, conjuntos de genes asociados a diferentes vías metabólicas, que permita determinar y predecir variaciones debidas al curso de una enfermedad.

Palabras Clave: ADN, Metagenómica, Estimación, Cluster, Clasificación

Contexto

Una línea de trabajo actual (Haegeman et al., 2013) considera que la evaluación de la biodiversidad requiere de un conjunto de índices de que representan riqueza, entropía etc. Esta perspectiva se amplía con la extrapolación de las curvas de rarefacción para evaluar la riqueza (Chao et al. 2014). Se propone entonces un método alternativo a las estimaciones no paramétricas de riqueza que habitualmente subestiman la cantidad de especies presentes en la comunidad. Respecto del microbioma humano, a escala global se encaran dos proyectos: el Metagenomics of the Human Intestinal Tract y el Human Microbiome Proyect . Pretenden analizar el rol del microbioma

en enfermedades tales como el cáncer, la obesidad, la inflamación intestinal o las patologías autoinmunes (Morgan y cols. 2013) En nuestro país el Plan Nacional de Ciencia, Tecnología e Innovación (Argentina Innovadora 2020), contempla el desarrollo de una Plataforma Tecnológica de Genómica y Bioinformática para estudios de ese tipo. El trabajo del grupo cuenta con la participación de biólogos investigadores de la Maestría en Explotación de Datos y Descubrimiento del Conocimiento de la UBA y de médicos pertenecientes al Instituto de Investigaciones Médicas Alfredo Lanari de la misma universidad.

Introducción

En biología computacional los componentes y procesos celulares son modelados por técnicas estadísticas y estudiados con recursos informáticos.

La metagenómica nació en 2004 al secuenciarse el ADN de los microorganismos presentes en aguas oceánicas. Luego los estudios se ampliaron a suelos y al microbioma humano.

Dentro del Proyecto C141 se ha desarrollado un algoritmo de recuento de especies para evaluar la riqueza microbiológica. Estimar este parámetro de biodiversidad resulta complejo pues se presentan problemas estadísticos. Se diseñó una metodología experimental para tratar la inferencia estadística de la riqueza poblacional y se supuso resueltos los efectos producidos por la secuenciación, el alineamiento, el filtrado y cualquier otro proceso del ADN obtenido a partir de una muestra de la comunidad. La medición de la riqueza microbiana requiere un gen marcador, de alta conservación a través de la evolución (Schloss-Handelsman 2006). Así, cuando las secuencias de individuos de una muestra revelan cierto porcentaje de variaciones entre ellas, tales variaciones pueden atribuirse a la diferencia de especies y no a simples ocurrencias aleatorias.

Dado un conjunto de secuencias del gen marcador cabe inferir desde ella la riqueza de la comunidad. Lo usual es que el tamaño muestral no sea suficiente para la tarea y en los hechos la riqueza sea subestimada (Hughes et al., 2001) Esto se debe a la presencia de una proporción importante de especies o taxones raros, en términos estadísticos, lo que hace muy poco probable que se encuentren en la muestra individuos que pertenezcan a todas o casi todas ellas. Los individuos de especies raras son muy pocos en relación con los individuos de especies abundantes en la comunidad y, además, las especies raras son mucho más que las abundantes por lo cual el tamaño de la muestra debería ser muy grande para inferir un valor aproximadamente real de la riqueza. (Roesch, L. et al. 2007). Este problema se ha intentado resolver por estimadores no parámetricos como CHAO y ACE (Chao-Lee 1992) que si bien han mejorado las estimaciones no han dado por resuelta la inferencia. Se propone entonces un método alternativo al construir un proceso aleatorio que va incrementando en forma simulada el tamaño muestral utilizando una estimación de la probabilidad de que un próximo individuo que se agregue a la muestra corresponda a una especie nueva. (Santa Maria-Soria 2013).

El Algoritmo de Recuento de Especies (ARE) diseñado actualiza la cantidad de individuos y especies en cada iteración y va haciendo crecer el número de especies en simultáneo con la disminución hacia cero de la probabilidad de hallar una nueva especie. Las pruebas realizadas a partir de muestras de

poblaciones simuladas y reales permiten ver que la riqueza resulta así mejor apreciada. El modelado experimental de la comunidad se hace computacionalmente posible y razonable al optimizar los tiempos de ejecución y contar con capacidades de procesamiento en paralelo (López y cols. 2014)

El grupo ha comenzado a investigar también sobre secuencias del microbioma humano en presencia de cáncer de colon o de enfermedad de Crohn. Esto plantea desafíos informáticos pues se requiere la elección adecuada de técnicas ya desarrolladas, nuevos desarrollos o nuevas secuencias de cálculo utilizando algoritmos que permitan evidenciar las fluctuaciones biológicas y médicas de interés. (Knights y cols. 2011) El trabajo se propone analizar el desempeño de algoritmos de data mining en microbiomas relacionados con las enfermedades citadas. Los procedimientos de agrupamiento (clusters) y los de ensamble de árboles de decisión, presentan una serie importante de variantes que se pretenden estudiar con el objetivo de establecer un clasificador de enterotipos, conjuntos de genes asociados a diferentes vías metabólicas, que permita determinar y predecir variaciones debidas a edad, grupo o estadío de enfermedad.

Líneas de Investigación, Desarrollo e Innovación

Las líneas de trabajo consisten en la investigación, el desarrollo y la comparación de técnicas de minería de datos aplicadas sobre bases constituidas por secuencias de ADN correspondientes a una colección de microorganismos Se pretende dar con los procedimientos más aptos para revelar la biodiversidad de tales comunidades, la interacción recíproca entre grupos de microorganismos y hallar los métodos de clasificación adecuados para la apreciación de las propiedades de la población de microbiana. Se trata de líneas multidisciplinarias donde los aspectos computacionales se deben considerar subordinados a la bondad de la caracterización biológica o de clínica médica que las técnicas elegidas aporten. Esto orienta la construcción de los modelos estadísticos y los algoritmos, así como también la elección de los recursos informáticos.

Resultados y Objetivos

Se realizaron dos tipos de pruebas con el algoritmo ARE. En la primera se construyó una población simulada con distribución de Fischer cuyo parámetros elegidos fueron $\alpha = 5000$ y x=0.995 (Magurran 2004). Se aplicó ARE sobre una muestra inicial de 1000 individuos y se comparó el desempeño de los estimadores CHAO, ACE y con la cantidad real de especies. La Tabla 1 evidencia las mejoras que ARE produce en la estimación de riqueza conforme aumenta el número de iteraciones.

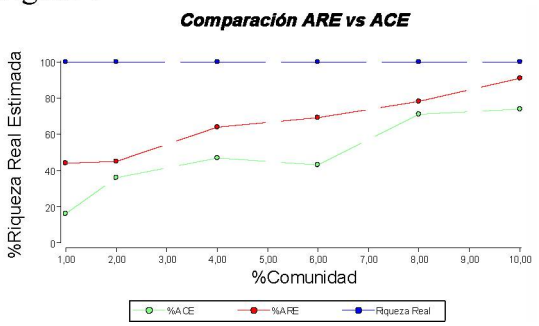
Tabla 1

Real	CHAO	ACE	ARE/ 45000	ARE/ 60000	ARE /200000	ARE/ 500000
26332	6699	6751	12821	14615	19271	24559
100%	25%	26%	49%	56%	73%	93%

Por otro lado se consideró la muestra de agua de mar profundo FS396,archea (Haegeman y cols. 2013) integrada por 16316 secuencias del gen 16S rRNA y cuya curva de rarefacción alcanza un comportamiento asintótico con ese tamaño. La cantidad de especies presentes en esta muestra, que se consideró como una comunidad completa, es 346. Se tomaron 5 muestras de cada uno de los tamaños 160, 320, 640, 960, 1280 y 1600 y sobre cada una se realizaron 10 corridas del algoritmo estableciéndose una estimación promedio. La Figura 1 muestra los resultados

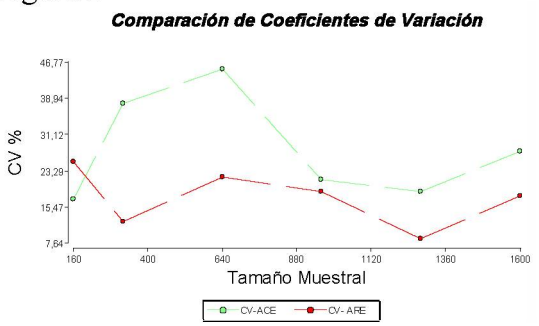
obtenidos según crece porcentualmente el tamaño muestral respecto al total poblacional. El trazo azul corresponde a la riqueza real, el rojo a la estimación ARE y el verde a ACE, que es el estimador no paramétrico habitualmente utilizado.

Figura 1



Además las pruebas arrojaron la menor variabilidad para la estimación ARE según se ve en la Figura 2.

Figura 2



Se concluyó entonces que el enfoque alternativo hace más precisa la estimación de la riqueza.

Con respecto al microbioma intestinal, para establecer las categorías de clasificación compatibles con la clínica médica existen dos enfoques. Uno por medio de la identificación de un solo gen marcador que apunta a evaluar características ecológicas como la riqueza y la diversidad del microbioma. El otro por el análisis completo del metagenoma, que agrupa las secuencias genéticas por su función metabólica y establece los enterotipos cuya distribución cambia en presencia de enfermedades.

El microbioma varía en cuanto a riqueza y distribución de abundancia de especies según distintas características pero no resulta tan variable cuando se consideran los grupos genéticos asociados con vías metabólicas al determinar los enterotipos. (Weinstock 2012) Es posible diferenciar una persona sana por la presencia de enterotipos asociados con la salud, de las personas enfermas cuyos enterotipos evidencian alteraciones metabólicas (Amiriam y cols. 2013). Las diferencias en la composición y la estructura del microbioma, aún en pacientes sanos, justifican la necesidad de afinar los métodos computacionales utilizados para reducir la probabilidad de error de clasificación y/o predicción. (Knights D y cols.. 2011). Los agrupamientos realizados por función genética permiten obtener categorías de clasificación. Luego se utilizan secuencias ya clasificadas en alguna categoría para entrenar un algoritmo que clasificará nuevas secuencias aún no categorizadas. Los árboles de decisión permiten minimizar el error de clasificación al ser entrenados y también testeados por secuencias ya categorizadas. Con estos métodos, dada una muestra del microbioma de un paciente, podría intentarse establecer si su diversidad, cantidad de especies y distribución, se corresponde con la presencia de una posible enfermedad. (Morgan y cols. 2012) Esto abre el camino para que la metagenómica sea una herramienta de análisis clínico (Statnikov y cols. 2013) al establecer la existencia de asociación entre la estructura y función del microbioma y la salud o enfermedad. Se han fijado dos objetivos.

- Evaluar la capacidad de algoritmos de clustering para realizar agrupamientos compatibles con los criterios clínicos sobre cáncer de colon y enfermedad de Crohn

- Analizar las variaciones de los agrupamientos del gen marcador causadas por motivos técnicos y efectuar un metanálisis estadístico con el clustering de enterotipos para determinar las categorías más estables de cada enfermedad analizada, compatibles con la clasificación clínica.

Formación de Recursos Humanos

En el equipo de trabajo participan un magister y un especialista en data mining, un doctor en biología, dos médicos, 2 ingenieros en sistemas, una matemática y un estudiante de ingeniería informática.

En años anteriores el trabajo ha dado lugar a una tesis de maestría y actualmente está en curso otra, además de un trabajo final de grado.

Referencias

Amiriam ES. Petrosino JF. Ajami NJ. Liu Y. Mims MP. Y Scheurer ME. (2013) Potential role of gastrointestinal microbiota composition in prostate cancer risk. Infectious Agents and Cancer. 2013 8:42

-Chao, A and Lee, S. 1992. Estimating the Number of Classes via Sample Coverage. Journal of American Statistical Association. Volume 87. Issue 417.

-Chao, A. Gotelli, N. J. Hsieh, T.C. Sander, E. L. Ma, K.H. Colwell, R. K. and Ellison A. M. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecological Monographs. 84 (1) 2014 pp.45-67.

-Haegeman, B. Hamelin, J. Motitarty, J. Neal, P. Dushoff, J. and Weitz, J. 2013 Robust estimation of microbial diversity in theory and in practice. ISME Journal 2013, 1-10

-Hughes, J. Hellmann, J. Ricketts, T. and Bohannan, B. 2001. Counting the uncounta-ble: statistical approaches to estimating microbial diversity. Applied and Environ-mental Microbiology. 4399-4406.

-López, L; Martínez, P; Cacho Mendoza, A; Soria, M y Santa María, C. (2014) Data Mining en Evaluaciones de Biodiversidad. XVI Workshop de Investigadores en Ciencias de la Computación. Pp 158-162.

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

-Magurran, A. 2004. Measuring Biological Diversity. Blackwell Science Ltd.

Morgan XC. Segata N. y Curtis Huttenhower. (2013) Biodiversity and functional genomics in the human microbiome. Trends in Genetics. Vol 29 No. 1

-Roesch, L. Fulthorpe, R. Riva, A. Casella, G. Hadwin, A. Kent, A. Daroub, S. Camargo, F. Farmerie, W. and Triplett, E. 2007. Pyrosequencing enumerates and con-trasts soil microbial diversity. The ISME Journal. 1, 283-290.

-Schloss, P. and Handelsman,J. 2006. Toward a census of bacteria in soil. PLoS Computational Biology. Volume 2.

- Santa María C. y Soria M. (2013) Inferencia de Parámetros de Biodiversidad por medio de Simulación" MACI Vol 4 1: 5-8

-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

-Weinstock GM. (2012) Genomic approaches to studyng the human microbiota. Nature Vol 489 250-256