APLICACIONES DE DATA MINING AL ESTUDIO DE LA BIODIVERSIDAD

Cristóbal R. Santa María Departamento de Ingeniería UNLAM
Marcelo Soria Facultad de Agronomía Cátedra de Microbiología UBA
Florencio Varela 1903 San Justo Pcia. de Buenos Aires
54-011-44808952
csantamaria@unlam.edu.ar
soria@agro.uba.ar

RESUMEN

El trabajo propone la utilización conjunta de técnicas de data mining y simulación para evaluar la riqueza y diversidad de comunidades microbianas. Se parte de una muestra formada por distintas secuencias de ADN que se alinean para luego ser agrupadas según su similaridad en clusters. Cada uno de estos clusters es una especie y el propósito es estimar su número y distribución en la comunidad basándose en la información que da la muestra. La técnica de rarefacción, sustentada en el procedimiento bootstrap, permite construir una curva cuya tendencia asintótica es precisamente la riqueza de la comunidad. Para alcanzar tal asíntota, y a la vez para estimar la distribución estadística de las especies, se propone una simulación que utiliza la estimación de Turing sobre la probabilidad de nueva especie al seleccionar un individuo nuevo y la idea de cobertura para la porción de la distribución que cubre la muestra.

Palabras Clave: Cluster-Riqueza-Diversidad-Rarefacción-Simulación-Cobertura

CONTEXTO

La línea de investigación que se presenta está inserta en el proyecto Aplicaciones de Data Mining al Estudio de la Biodiversidad en Relevamientos Metagenómicos que, dentro del marco del Programa de Incentivos a la Investigación, se lleva adelante en del Departamento de Ingeniería e Investigaciones Tecnológicas de la UNLAM. Tal tarea se realiza con la colaboración de un investigador de la Cátedra de Microbiología de la Facultad de Agronomía de la UBA y con el asesoramiento de la Maestría en Explotación de Datos y Descubrimiento del Conocimiento de la Facultad de Ciencias Exactas y Naturales de la UBA.

INTRODUCCIÓN

La metagenómica consiste en el relevamiento de las comunidades microbianas de un ecosistema, como pueden ser la flora intestinal, los microorganismos asociados al suelo o a diferentes cuerpos de agua, etc. El relevamiento se realiza a partir de una muestra compuesta del material genético de los miembros que conforman esas comunidades. Mediante nuevas tecnologías de secuenciación de ADN se pueden obtener cientos de miles y hasta millones de secuencias al mismo tiempo. Esto plantea importantes desafíos estadísticos y abre la perspectiva para la aplicación de técnicas de Data Ming y KDD en forma creciente. Ver Guazzaroni et al [1]

El objetivo es analizar la biodiversidad de la comunidad que se determina a través de la riqueza de especies y de su distribución relativa. Ver Magurran [2]. El análisis comparativo de estos componentes de la biodiversidad en comunidades microbianas de suelos es una herramienta de diagnóstico sensible para evaluar la calidad de los suelos y la sustentabilidad de las prácticas agrícolas. Ver los trabajos de Parks y

Beiko [3], Raes et al. [4] y Hollister et al [5].

En los estudios de diversidad biológica de microorganismos la riqueza, expresada como número total de especies, es desconocida y sus estimaciones varían en órdenes de magnitud. Tampoco existen hipótesis adecuadas o debidamente probadas acerca de la distribución del número de individuos por grupo. Ver Schloss y Handelsman [6], y Hill et al[7]

A partir de una muestra del material se extraen y purifican fragmentos de ADN correspondientes a los microorganismos presentes en la comunidad. El ADN purificado se procesa en secuenciadores que permiten construir sucesiones de símbolos de las componentes químicas que lo integran. Esta es la información que puede almacenarse y procesarse por medio de computadoras.

La gran cantidad de datos asociados a un estudio metagenómico y el relativo desconocimiento sobre la estructura de las comunidades microbianas, y por ende, la dificultad de establecer procedimientos de análisis basados en modelos, sugieren abrir paso a metodologías de data mining. Ver Roesch et al [8] y Klimke et al[9]. La idea del trabajo es analizar aspectos del empleo de técnicas de data mining, originadas en la estadística, para cuantificar la biodiversidad.

El ADN tiene una molécula lineal extremadamente larga, un polímero, compuesto por la sucesión de cuatro componentes modulares o monómeros: adenina (A), citosina (C), guanina (G) y timina (T). El ADN total de un microorganismo, también llamado genoma, puede abarcar unos cuatro millones de nucleótidos, en una cadena circular.

La metagenómica realiza el análisis genómico de comunidades microbianas al combinar el material genético de distintos organismos del medio. Ver Liza Gross [10]. En este trabajo se utilizará el "análisis de marcadores". El objetivo es obtener secuencias de uno, o unos pocos, genes predeterminados, para establecer a qué especie pertenecen. En particular se utilizará el gen que codifica para el ARN ribosomal de 16S (16S rRNA), cuyo uso en estimaciones comparativas de riqueza puede verse por ejemplo en los trabajos de White et al. [11] y Youssef y Elshahed [12]. Además este gen es ampliamente utilizado en estudios filogenéticos que buscan dilucidar la evolución y el desarrollo de las especies. Se comparan secuencias y se trabaja con las diferencias entre éstas para estructurar árboles filogenéticos y sucesiones evolutivas. Ver Brady y Salzberg [13]

Las relaciones evolutivas entre los organismos se pueden representar mediante árboles filogenéticos. Cada árbol supone un ancestro común a todas las especies; es decir una expresión raíz para la cadena del 16S rRNA que se modifica al avanzar hacia las hojas. Los agrupamientos se realizan utilizando distintos niveles de similaridad o disimilaridad que se asocian con categorías taxonómicas tales como especie, género, familia, orden, clase, phylum y dominio. Los grupos que se constituyen en cada nivel de disimilaridad sirven para evaluar la diversidad en cada categoría taxonómica y se los denomina Unidades Taxonómicas Operacionales (OTUs). Ver Schloss y Handelsman [14]. El primer paso es alinear las distintas secuencias y definir una "distancia genética" que, una vez alineadas, permita representar la diferencia entre dos cadenas de ADN. En la Figura 1 se muestra un ejemplo de alineamiento.

Figura 1

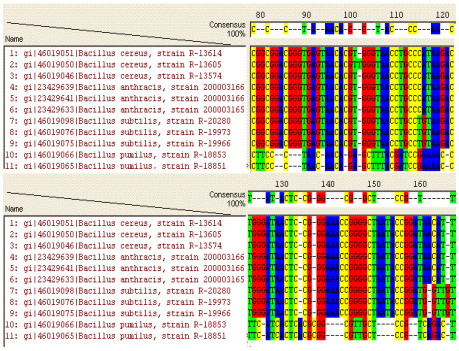
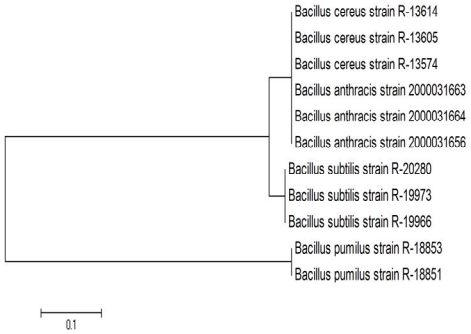


Figura 2



En su forma más simple la disimilaridad entre dos secuencias es igual al número de posiciones enfrentadas por el alineamiento que presentan distinto contenido. Por lo tanto una idea es usar la distancia de Hamming que se calcula como la proporción de celdas en las que la diferencia entre residuos ocurre. Sin embargo hay algunas cuestiones de naturaleza biológica que sugieren cambios en el cálculo de las distancias a través de distintos modelos expuestos por Hillis et al. en [15].

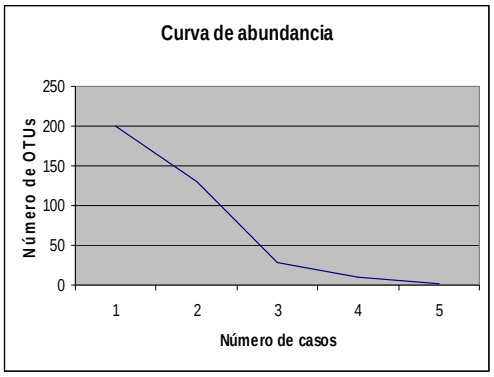
Se calcula luego la matriz de distancias utilizando todas las secuencias disponibles. Cada celda de la matriz es la distancia que hay entre la secuencia de la fila y la de la columna. A continuación se fija el criterio con el que las secuencias se considerarán similares o disimilares. Los distintos porcentajes de disimilaridad indican la máxima diferencia que se acepta entre cadenas de 16S rRNA correspondientes a los individuos de un grupo. Es de práctica considerar que una disimilaridad de hasta el 3% corresponde a individuos de la misma especie mientras que para una disimilaridad que no exceda el 5% se considera igual género o que, para otra del 20%, hay igual clase o phylum. El árbol resulta un dendrograma como se ve en la Figura 2 y se corta al nivel del taxón requerido por el estudio

La construcción de las OTUs puede realizarse de distintas formas. El criterio del “vecino más cercano” asigna una secuencia en una OTU cuando su disimilaridad con una cualquiera del grupo no supera el porcentual elegido. En cambio el método del “vecino más lejano” es más restrictivo pues asigna una secuencia en una OTU cuando su disimilaridad con cada una de las secuencias que la integran no supera el porcentaje definido. El procedimiento del “vecino promedio” utiliza la disimilaridad promedio entre las secuencias de una OTU y las que están fuera de él para asignar una secuencia cuando su disimilaridad con las del grupo resulta menor que dicho promedio. Ver Everitt [16]

Para medir la diversidad de una comunidad biológica se tienen en cuenta diferentes conceptos: riqueza, abundancia, uniformidad y dominancia. Ver Magurran [2]. Las mediciones son estimaciones estadísticas que tratan de atenuar la incertidumbre producida por el gran volumen de datos y la incerteza sobre la cantidad de especies presentes. Los diferentes índices de riqueza estiman el número de especies distintas mientras que los modelos de abundancia establecen la distribución del número de veces que se presentan las especies en la comunidad. Ver Hill et al. [7]. En la

Figura 3 se ve un ejemplo de curva de abundancia

Figura 3.



A efecto del modelado la abundancia puede presentarse también ordenando las frecuencias de aparición de cada OTU de mayor a menor $X_1, X_2, ..., X_S$. En la medida que el tamaño de la muestra crezca tales frecuencias tenderán a adoptar el valor de probabilidad de aparición de una OTU en una muestra de la población. La probabilidad de aparición se puede modelar siguiendo distintas distribuciones. La más usada suele ser la distribución geométrica según expone Ann Chao en [17]. Se tiene: $p_i = \alpha(1-\alpha)^{i-1}$ con $i = 1,...,S$ y α un parámetro.

Por lo tanto la distribución de las especies en la comunidad podría tener diversas formas según los valores de p_i. En un extremo podría suponerse que en la comunidad hay la misma cantidad de organismos de cada especie. En el otro, la suposición sería que todos los organismos pertenecen a una misma y única especie. Se diría entonces que en el primer caso hay uniformidad de especies mientras que en el segundo hay dominancia de la única especie. El grado de la relación uniformidad-dominancia quedará establecido por algunas especies que aparezcan comúnmente y por otras que resulten raras. Ver Magurran [2]. Al tomar una muestra aleatoria de la comunidad es posible entonces que ciertas especies que debieran ser contabilizadas en su riqueza, no aparezcan por ser raras o al menos no las más comunes. Si se trata de microorganismos: ¿cuál debe ser el tamaño de la muestra para asegurar la presencia en ella de todas o casi todas las especies? La respuesta no solo depende de la relación uniformidad-dominancia desconocida "a priori" sino también de las cantidades de especies posibles y de organismos en la comunidad que tampoco se conocen. Dos medidas comúnmente usadas para medir riqueza y diversidad son el índice S de riqueza de especies y la medida E de uniformidad de especies.

El índice de riqueza S es el número de especies que hay en el medio. Su estimación puede hacerse utilizando curvas de rarefacción que se basan en el procedimiento de remuestreo bootstrap usual en data mining. Ver Efron [18]. El índice de uniformidad E utiliza el concepto de entropía definido en la teoría de la información construida por Shannon en [19]. La entropía se calcula de acuerdo a $H = -\sum p_i \ln p_i$ dónde p_i es la probabilidad de ocurrencia de la i-ésima especie. Si hay uniformidad de especies, es decir si todas las especies tienen la misma probabilidad de ser observadas la entropía es máxima. De tal forma para una cantidad $S = n$ de especies resultaría

$$H = -\sum_{i=1}^{n} \frac{1}{n} \ln \frac{1}{n} = -n \frac{1}{n} \ln \frac{1}{n} = -\ln \frac{1}{n}$$

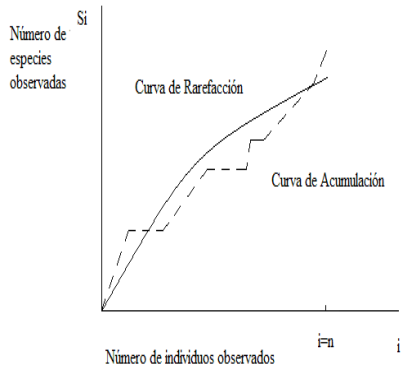
Para definir un índice normalizado se toma $E = \frac{H}{\ln S}$ y en este caso, si hay uniformidad, queda

$$E = \frac{-\ln \frac{1}{n}}{\ln n} = \frac{-\ln \frac{1}{n}}{\ln\left(\frac{1}{n}\right)^{-1}} = \frac{-\ln \frac{1}{n}}{-\ln \frac{1}{n}} = 1$$

Si la distribución de especies en la población se va deslizando desde la uniformidad hacia la dominancia el índice E se moverá consecuentemente hacia 0. En efecto; en la situación teórica extrema en que exista una sola especie resulta $H = -p_{especie} \ln p_{especie} = -1 \ln 1 = 0$ Y entonces también $E = 0$

La idea básica de la rarefacción es que a partir de la toma de muestras de mayor tamaño será posible capturar un número creciente de especies distintas. Dada una comunidad que tenga una cantidad desconocida N de individuos y un número S de especies distintas también desconocido, se pueden tomar muestras de tamaño n y determinar S_n que es el número de especies distintas halladas en una muestra. El valor esperado teórico $E(S_n)$ se aproxima por el promedio de los S_n y se utiliza para medir la riqueza S del medio. Se comienza construyendo una curva de acumulación del número de especies distintas según se ve en la gráfica punteada de la Figura 4.

Figura 4



La curva punteada muestra la acumulación del número de especies distintas conforme se van examinando cada uno de los n individuos que forman la muestra. El procedimiento de rarefacción consiste en repetir muchas veces este examen tomando cada vez un orden distinto y aleatorio de los n casos, y estableciendo el número acumulado promedio para cada cantidad i de casos examinados. La curva resultante tiene un aspecto suave como se observa en la línea llena de la Figura 4. Es decir; la curva de rarefacción representa el promedio de todas las curvas de acumulación construidas. Ver Gotelli y Colwell [20]. Si se llama S_{iobs} al valor de ordenadas de la curva de rarefacción empíricamente construida en cada valor i, pueden establecerse intervalos de confianza para S_i. Hughes y Hellmann [21]. De tal modo se obtiene, para un tamaño n de la muestra, una estimación S_n de la riqueza de la comunidad y al aumentar n es posible hallar una asíntota horizontal cuyo valor resulte una aproximación a la riqueza del medio. Si se considera que el tamaño de la muestra tomada no es suficiente para determinar la tendencia asintótica se puede recurrir a la simulación del muestreo sobre la base de la distribución de frecuencias de especies empírica.

LÍNEAS DE INVESTIGACIÓN Y DESARROLLO

Se trabaja actualmente en la aplicación de técnicas de data mining a la determinación de la riqueza y diversidad de las comunidades microbiológicas. Se consideran datos extraídos de suelos disponibles en la base de datos de NCBI (National Center for Biotechnology Information) a los que se accede on-line en http://www.ncbi.nlm.nih.gov/ Estos datos se procesan con el software libre MOTHUR disponible en http://www.mothur.org/. De esta manera se leen las secuencias que se alinean y filtran. A continuación con el programa DNADIST de la Suite PHYLIP, disponible en

http://evolution.genetics.washington.edu/phylip.html, se calculan las distancias entre las distintas secuencias y se arma la matriz de distancias. Luego con el programa Cluster de MOTHUR se realizan los agrupamientos para el nivel de disimilaridad elegido y con el programa Rarefaction del mismo software se construyen las curvas de rarefacción de las muestras utilizadas. Como dado el tamaño de las muestras utilizadas, que es el que habitualmente puede recogerse en este tipo de estudios, las curvas no logran establecer una tendencia asintótica que estime la riqueza del medio, se realiza una simulación basada en la distribución empírica de frecuencias de especies obtenidas. Con ello se incrementa el tamaño de la muestra al buscar encontrar el comportamiento asintótico de la curva de rarefacción. Al agregar un nuevo individuo éste puede resultar perteneciente a una especie ya conocida o no. La sucesión de los valores de cantidad de especies resulta entonces un proceso aleatorio como se describe a continuación. En el Diagrama 1 la variable i representa el número de individuos considerado en la muestra y S_i es la cantidad de especies halladas cuando la muestra tiene tamaño i.

Diagrama 1

…	S_{i-2}	S_{i-1}	S_i	S_{i+1}	S_{i+2}	…
…	i-2	i-1	i	i+1	i+2	…

Para estimar la probabilidad de que el i-ésimo individuo corresponda a una especie nueva se toma el estimador de Turing (Ver Good [22]) $\hat{f}_0 = \frac{n°sgletones}{i-1}$, donde cada singletón en la muestra de i-1 individuos está formado por el solo individuo que representa a una especie en esa muestra. Resulta entonces que para cada valor de S_i hay una probabilidad asociada como se ve en la Tabla 1.

Tabla 1

Estado	Probabilidad estimada
$S_i=S_{i-1}$	$p = 1 - \hat{f}_0$
$S_i=S_{i-1}+1$	$p = \hat{f}_0$

Así el valor esperado de S_i, que correspondería a la curva de rarefacción para i individuos considerados se calcula:

$E(S_i) = S_{i-1}(1-\hat{f}_0) + (S_{i.-1}+1)\hat{f}_0$ y operando se obtiene $E(S_i) = S_{i-1} + \hat{f}_0$

Si se realiza una simulación eligiendo de a uno individuos en una muestra, cabría esperar que cuando i crezca $\hat{f}_0 \to 0$ pues el número de especies aún no encontradas debiera ir disminuyendo al ser finita la cantidad S de especies buscada. Es decir, en la realidad así debiera ocurrir por lo cual es necesario que el modelo matemático adoptado lo tenga en cuenta. Esto se puede lograr apelando al concepto de cobertura del muestreo según lo presentan Chao y Shen ([23]). En el modelo que utilizan se supone que cada una de las S especies existentes en el medio tiene una probabilidad π_j de aparición. Si se toma una muestra de tamaño i, de forma tal que a cada especie le correspondan x_j individuos de la misma, se define la cobertura como $C = \sum_{j=1}^{S} \pi_j I[x_j > 0]$ donde I es la función indicador que vale 1 si $x_j > 0$ y 0 en otro caso. El valor de S es desconocido y en realidad en la expresión de C solo suman aquellas especies que efectivamente aparecen. Claramente $0 \le C \le 1$. Si $C = 0$ es porque no ha aparecido aún ninguna especie (caso solo teórico e imposible si se tomó una muestra) y si $C = 1$ es porque todas las especies existentes han aparecido en la muestra. Además a partir de la muestra puede calcularse el número de especies representadas por m

individuos $g_m = \sum_{j=1}^{S} I(x_j = m)$

suponiendo una cantidad g_0 que sea precisamente el número de especies con 0 individuos. Obsérvese que puede ocurrir que $g_m = 0$ para varios valores de $m = 1,...,i$. Así $\sum_{m=1}^{i} g_m = S_i$, donde S_i es la cantidad de especies halladas en la muestra que tiene i individuos y claramente resultaría $S = S_i + g_0$. Además g_1 es el número de singletones en la muestra y $\sum_{m=1}^{i} m g_m = i$ es el tamaño muestral.

Según exponen Chao y Shen [23], un estimador de la cobertura según la muestra tomada es $\hat{C} = 1 - \frac{g_1}{i} = 1 - \hat{f}_0$ y la probabilidad π_j de elección de un individuo de la j-ésima especie se estima por $\hat{\pi}_j = \frac{x_j}{i}\hat{C}$

La idea de cobertura muestral es utilizada por Chao y Lee para construir un estimador no paramétrico de la riqueza de especies en [24]. Con similar criterio se puede sugerir entonces la corrección del estimador $\hat{f}_0$ de tal manera que, en la medida en que se incrementa la cobertura, la probabilidad de especie nueva estimada disminuya al restar la aparición de menos especies.

La programación de tales alternativas planteadas se realiza utilizando el software libre R.

RESULTADOS Y OBJETIVOS

Las necesidades de la investigación y desarrollo en el campo biológico respecto de contar con estimaciones fidedignas de riqueza y diversidad en comunidades microbianas, en especial de suelos, no siempre se ven satisfechas dadas las muy diferentes respuestas que arrojan los distintos modelos empleados para medirlas. El trabajo se propone utilizar las técnicas de data mining combinándolas con la simulación de casos para lograr un mejor desempeño en la predicción de dichas cantidades. En particular se espera llegar a verificar que los procedimientos de rarefacción y simulación combinados mejoran la exactitud de las mediciones de biodiversidad realizadas sobre las comunidades.

Hasta el momento se han logrado resultados parciales basados en el proceso del conjunto de datos de la transecta SRA009427 obtenida de NCBI (ver [5]). Cada conjunto de secuencias fue alineado, filtrado, se calculó la matriz de distancias y realizó el agrupamiento según el nivel de disimilaridad elegido del 3%. A continuación se construyó en cada caso la curva de rarefacción. Lo expuesto se resume mostrando el ejemplo de proceso del subconjunto de datos SRR030385 en la Tabla 2 y la Figura 5

Tabla 2

n tamaño muestral n=1641

label	sobs	chao	shannon
0.05	542	821.181034	5.782979

Sobs corresponde a la riqueza medida por rarefacción. La cantidad chao es el valor de un índice de origen no paramétrico que se utiliza como modelo alternativo para medir la riqueza. shanon es la entropía resultante. IncNU es la proporción del incremento de especies respecto del mínimo minS que se alcanzaría si la distribución de las mismas fuese uniforme. Su valor mínimo es 0 cuando la distribución es uniforme

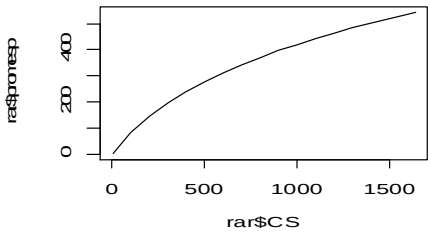
minS=324.7251071

IncNU=0.604563313

Singletones=255

p probabilidad de descubrir nuevas especies (estimación de Turing)
p=255/1641= 0.155393053
$promesp es la cantidad promedio de especies según los tamaños muestrales

Figura 5



Análogamente se consideraron los subconjuntos SRR030389 al SRR030392 que constituyen muestras diferentes. En todos los casos se observó que la curva de rarefacción no llegó a alcanzar el comportamiento asintótico requerido para estimar la riqueza adecuadamente. Esto reforzó la idea inicial de utilizar un muestreo simulado basado en la distribución de frecuencias de las especies en las muestras de tamaño n.

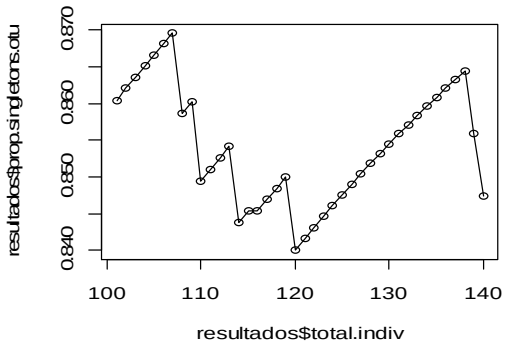
Consecuentemente se desarrolló un programa en lenguaje R que utilizando la estimación de Turing incorporase individuos a la muestra. Para una submuestra de 100 individuos y luego del proceso de alineado, filtrado y armado de la matriz de distancias se obtuvo la distribución inicial de frecuencias que se ve en la Tabla 3.

Tabla 3

dens	n.otus	tot.indv	freq.indv
1	73	73	0.85882353
2	9	18	0.10588235
3	3	9	0.03529412

A partir de aquí, se simuló la incorporación de individuos a la muestra observándose que el estimador de Turing no convergía rápidamente a 0 según se observa en la Figura 6.

Figura 6



Esto motivó la necesidad de incorporar la idea de cobertura cuya programación se desarrolla actualmente.

FORMACION DE RECURSOS HUMANOS

El equipo de trabajo está formado por el Dr. Marcelo Soria, biólogo dedicado a la investigación en bioinformática y docente en la Maestría en Explotación de Datos y Descubrimiento del Conocimiento de la Facultad de Ciencias Exactas y Naturales de la UBA y por el Especialista Cristóbal R. Santa María, matemático, investigador y docente en el Departamento de Ingeniería de la UNLAM quien obtuvo en 2008 la especialización en data mining como parte de la Maestría antes citada y se encuentra preparando su tesis para obtener la misma.

REFERENCIAS

[1] [2009] Guazzaroni, M.E, Beloqui, A, Golyshin, P y Ferrer, M. “Metagenomics as a new technological tool to gain scientific knowledge”. World Journal Microbiologic Biotechnology. 25:945-954

[2] [2004] Magurran, A. Measuring Biological Diversity. Blackwell Science Ltd

[3] [2010] Parks, D y Beiko, R. “Identifying biologically relevant differences between metagenomic

communities". Bioinformatics Advance Access published February 3. The Oxford University Press.
[4] [2007] Raes, J, Foerstner, K. U y Bork, P. "Get the most out your metagenome: computacional analysis of environmental sequence data". Current Opinion in Microbiology. 10:490-498
[5] [2010] Hollister, E, Engledow, A, Hammett, A, Provin, T, Wilkinson, H. y Gentry, T. "Shifts in microbial community structure along an ecological gradient of hypersaline soils and sediments". The ISME Journal. 1-10.
[6] [2006] Schloss, P. y Handelsman,J. "Toward a census of bacteria in soil". PLoS Computational Biology. Volume 2.
[7] [2003] Hill, T, Walsh, K, Harris, J yMoffett, B. "Using Ecological Diversity Measures with Bacterials Communities". FEMS.Microbiology Ecology 43 1-11
[8] [2007] Roesch, L, Fulthorpe, R, Riva, A, Casella, G, Hadwin, A, Kent, A, Daroub, S, Camargo, F, Farmerie, W y Triplett, E. "Pyrosequencing enumerates and contrasts soil microbial diversity". The ISME Journal. 1, 283-290.
[9] [2009] Klimke, W, Agarwala, R, Badretdin, A, Chetvernin, S, Ciufo, S, Fedorov, B, Kiryutin, B, O'Neill, K, Resch, W, Resenchuk, S, Schafer, S, Tolstoy, I y Tatusova, T. "The national center for biotechnology information's proteins clusters database". Nucleic Acids Research. Vol. 37.
[10] [2007] Gross, L. "Untapped Bounty: Sampling the Seas to Survey Microbial Biodiversity". PLoS Biology/ Volume 5/Issue 3/e85
[11] [2009] White, J, Nagarajan, N, Pop, M. "Statistical methods for detecting differentially abundant features in clinical metagenomic samples". PLoS Computational Biology. Volume 5.
[12] [2008] Youssef, N y Elshahed, M. "Species richness in soil bacterial communities: A proposed approach to overcome sample size bias". Journal of Microbiological Methods. 75 86-91.
[13] [2009] Brady, A y Salzberg, S. "Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models". Nature Methods. Vol. 6 N° 9.
[14] [2005] Schloss, P y Handelsman, J. " Introducing DOTUR, a Computer Program for Defining Operational Taxonomic Units and Estimating Species Richness". Applies and Environmental Microbiology. Pgs. 1501-1506
[15] [1996] D.M. Hillis, C. Moritz, B.K. Mable. Molecular Systematics. Second Edition, Sinauer Associates, Inc. Publishers. Sunderland, MA. USA
[16] [2001] Everitt, B, Landau, S, Leese, M. Cluster Analysis. Fourth Edition. Arnold.
[17] [2003] Chao, A. "Species Richness Estimation". Technical Report. Institute of Statistics. National Tsing Hua University.
[18] [1978] Efron, B. "Computers and theory of statistics: thinking the unthinkable". Technical Report N° 39. Division of Biostatistics. Stanford University
[19] [1949]. Shannon, C. The mathematical theory of communication. Illini Books edition. 1963.
[20] [2001] Gotelli, N y Colwell,R. "Quantifying biodiversity: procedures and pitfalls in measurement and comparison of species richness" Ecology Letters. 4: 379-391
[21] [2005] Hughes,J y Hellmann, J. "The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity". Methods in Enzymology. Vol 397.
[22] [1953] Good, I. "The population Frequencies of Species and the Estimation of Population Parameters". Biometrika. Vol 40 N| ¾ pp 237-264

[23] [2003] Chao, A y Shen, T. “Nonparametric estimation of Shannon’s index of diversity when there are unseen species in sample”. Environmental and Ecological Statistics 10, 429-443.
[24] [1992] Chao, A y Lee, S. “Estimating the Number of Classes via Sample Coverage”. Journal of American Statistical Association. Volume 87. Issue 417.