Estimación de Biodiversidad por Data Mining y Simulación

Cristóbal Santa María[1] y Marcelo Soria[2]

[1]Universidad Nacional de La Matanza. Departamento de Ingeniería.
smaria@sion.com
[2]Universidad de Buenos Aires. Facultad de Agronomía. Cátedra de Microbiología.
soria@agro.uba.ar

Resúmen. Los estudios de diversidad microbiológica basados en información genética involucran procesos computacionales fundados en la minería de datos y en la estimación estadística. Con el objetivo de obtener la riqueza de especies, entendida como el número de las mismas, y su distribución en el medio, se procesan secuencias del gen 16S rRNA. Se mide la "distancia genética" entre secuencias y se procede a un agrupamiento en "clusters" a partir del cual se realiza el recuento. Las estimaciones estadísticas estándar tropiezan con la incertidumbre creada por la insuficiencia del tamaño muestral lo que requiere explorar formas alternativas de estimación. El trabajo presenta dos formas de evaluación de la riqueza a partir de una estimación de la probabilidad de existencia de especie nueva y del concepto de entropía. A tal efecto se utiliza la simulación como una técnica de explotación de los datos muestrales con la que se obtienen resultados que mejoran las estimaciones habituales.

Palabras Claves: Biodiversidad-Riqueza-Estimación-Simulación.

1 Introducción

En forma general, puede decirse que, para evaluar la biodiversidad de un medio, hay que calcular las cantidades de taxones, ya sean especies, familias u otros, a efecto de establecer su riqueza y analizar además la forma en se distribuyen.

La tarea requiere acordar, en primer lugar, un criterio biológico para identificar los taxones. Una alternativa crecientemente utilizada al respecto es el análisis basado en el gen 16S rRNA que ha tenido una alta conservación a lo largo del proceso evolutivo y que permite, por ello, apreciar con exactitud las diferencias taxonómicas [1]. Una vez secuenciadas las cadenas de ADN del gen se las alinea y se miden las "distancias genéticas" para realizar un agrupamiento en "clusters". Los distintos umbrales de disimilaridad que se eligen para formar estos grupos permiten establecer el nivel taxonómico, especie o familia por ejemplo

Pero, cuando se desea inferir desde una muestra la riqueza de todo el medio biológico, se presentan además otras dificultades, de carácter estadístico, que provienen de la gran cantidad de microorganismos que integran realmente la comunidad y de la existencia de taxones que se encuentran en muy baja proporción y resultan, por ende, raros. Ocurre entonces que el tamaño de la población y la rareza estadística de algunos taxones, cuya importancia biológica puede ser mucho más significativa que su número, se suman a limitaciones tecnológicas y/o económicas para introducir un grado de incertidumbre en las estimaciones de biodiversidad

poblacional a partir de muestras, que no puede tratarse con las técnicas estadísticas habituales.

Los modelos en uso suelen subestimar la real cantidad de taxones presentes en la comunidad y, por lo tanto, desconocer una parte de la distribución de los mismos [2]. La idea del presente trabajo es explorar alternativas para aportar en las determinaciones de biodiversidad por la aplicación de técnicas de minería de datos y simulación estadística. El trabajo es continuidad del presentado en WICC2011 señalado en [3] y se realiza en el marco de la preparación de una tesis de maestría en Explotación de Datos y Descubrimiento del Conocimiento en el Departamento de Computación de la Universidad de Buenos Aires

2 Contexto

Una contextualización que abarca la forma computacional del ADN, el concepto metagenómico, la construcción del árbol filogenético y la "clusterización" de las secuencias genéticas, fue desarrollada en [3]

La diversidad biológica de una comunidad es función conjunta de dos conceptos: la riqueza, entendida como la cantidad de especies (o de taxones en general) y la distribución de las especies (o taxones) en su sentido estadístico [4]. Los modelos matemáticos utilizados [5] conducen a distintas estimaciones entre las que se focaliza la riqueza como principal parámetro poblacional. Ésta debe inferirse a partir de muestras. La distribución en el medio puede presentar algunas especies dominantes, otras que no lo son tanto y una mayoría que en términos estadísticos resultan raras [4]. De modo que al elegir al azar un individuo de la comunidad, la probabilidad p_i de que sea de la especie i puede ser alta, media o muy baja según el grado de dominancia o rareza que tenga la especie. En una muestra de la comunidad se selecciona una cantidad x_i de individuos de esa especie y, suponiendo independencia en la elección, se tiene un vector $(x_1, x_2, ..., x_S)$ con las cantidades presentes de cada especie. La probabilidad de elegir esa muestra se distribuye en forma multinomial según:

$$p(x_1, x_2, ..., x_S) = \binom{n}{x_1, x_2, ..., x_S} p_1^{x_1} p_2^{x_2} \dots p_S^{x_S} \quad (2.1)$$

donde S es la cantidad de especies

$$n = x_1 + x_2 + \dots + x_S \quad (2.2)$$

y

$$\binom{n}{x_1, x_2, ..., x_S} = \frac{n!}{x_1! x_2! \dots x_S!} \quad (2.3)$$

Ver [6]

Cuándo para la muestra escogida, algunos valores de x_i son cero, esto significa que no se va a tener registro de la presencia de esas especies en la comunidad. Resulta entonces que se estiman menos especies de las que realmente

hay. En particular es claro que una especie rara tiene menos probabilidad de figurar en la muestra que una dominante. La respuesta tradicional de la estadística sería aumentar el tamaño de la muestra Pero esto no es posible ni razonable.

Ann Chao desarrolló en [7] un nuevo estimador de la riqueza de especies.

$$S_{Chao\ 1} = D + \frac{f_1^{\ 2}}{2(f_2 + 1)} - \frac{f_1 f_2}{2(f_2 + 1)^2} \tag{2.4}$$

dónde f_r es la cantidad de especies que aparecen r veces en la muestra de tamaño

$$n = \sum_{r=1}^{n} r f_r \tag{2.5}$$

$$D = \sum_{r=1}^{\infty} f_r \tag{2.6}$$

es la cantidad total de especies observadas y claramente si N es la cantidad desconocida de especies en el medio resulta

$$D = N - f_0 \tag{2.7}$$

con f_0 número de especies que no aparecen en la muestra también desconocido. El índice CHAO funciona bien como estimador siempre que sea considerada una cota inferior [12].

Con el fin de mejorar la estimación en [8] se construye un nuevo estimador basado en la idea de cobertura de la muestra.

La cobertura se obtiene como suma de las probabilidades de las especies. Debe crecer cuando mas especies están en la muestra, pero lo hace en relación con la proporción que cada especie va teniendo. La idea es que también tiene en cuenta las variaciones de la probabilidad además de la porción de la distribución cubierta. Su estimación se realiza por medio de:

$$\hat{C} = 1 - \frac{f_1}{n} \tag{2.8}$$

donde f_1 es el número de especies representadas en la muestra por un solo individuo. La fracción

$$T = \frac{f_1}{n} \tag{2.9}$$

Es una cantidad sugerida por Turing [9] para estimar la probabilidad de descubrir una especie nueva al agregar un nuevo individuo a la muestra. En [8] se utiliza además una estimación del coeficiente de variación γ de las probabilidades de las clases, lo da la fórmula del estimador de riqueza ACE (Abundance Coverage Estimator):

$$S_{ACE} \approx \frac{E(D)}{E(C)} + \frac{E(f_1)}{E(C)} \gamma^2 \tag{2.10}$$

Una alternativa a la estimación no paramétrica, son las curvas de rarefacción. El método permite estimar la riqueza de un medio aunque generalmente es aplicado para comparar riqueza entre dos o más comunidades pues alivia los problemas derivados del tamaño insuficiente y habitualmente desigual de las muestras [10]. 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 de $E(S_n)$ 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 conforme se van examinando cada uno de los n individuos que forman la muestra. El procedimiento de rarefacción aplica técnicas de remuestreo [11]. Consiste en repetir muchas veces este análisis de individuos, tomando cada vez un orden distinto y aleatorio de los n casos muestrales y estableciendo el número acumulado promedio para cada cantidad i de casos examinados. La curva de rarefacción resultante, representa el promedio de todas las curvas de acumulación construidas.

Si se llama S_{iobs} al valor de ordenadas de la curva de rarefacción en cada valor i, pueden establecerse intervalos de confianza para S_i. Este valor esperado resultaría el promedio S_i de las cantidades de especies, tomado sobre todos los órdenes posibles para cada valor de i, pero solo se consideran algunos de ellos para construir un intervalo de confianza que estima S_i. De tal modo se obtiene, para un tamaño n de la muestra, una estimación S_n.

El valor esperado $E(S_n)$ es un estimador asíntótico de S, la riqueza de la comunidad. De acuerdo a ellos podría incrementarse la cantidad de elementos en la muestra a fin de lograr una estimación adecuada. Por desgracia esto no es operativamente posible pues la cantidad de individuos necesarios para que la muestra sea representativa es desconocida y seguramente muy grande para las posibilidades de la tecnología de secuenciación y de proceso de los datos [12]. Sin embargo si se utilizan curvas de rarefacción para evaluar la riqueza es posible establecer una asíntota horizontal cuyo valor resulte al menos una aproximación a la riqueza del medio. Se construye para esto una expresión analítica de la curva que permita extrapolar su comportamiento En particular puede realizarse un ajuste de mínimos cuadrados utilizando una hipérbola rectangular [13]. Como un camino alternativo se cita en [14] el uso de una curva exponencial. Una vez hallada la expresión matemática, se extrapola el comportamiento asintótico y se halla así un valor de $\hat{S}_n$ que aproxime suficientemente a S, la riqueza de la comunidad.

Varias curvas de rarefacción calculadas para muestras de distinto tamaño permiten comparaciones de riqueza según el punto a partir del cual se considera que se alcanzó el valor de la asíntota horizontal. Pero a la vez, la rapidez

con que esa asíntota sea alcanzada en cada caso, puede permitir una apreciación de la diversidad.

La aplicación de las distintas formas de estimación detalladas resulta generalmente en una subestimación de los verdaderos valores de riqueza según [15]. Por este motivo se plantea la necesidad de mejorar las técnicas de estimación que se emplean.

3 Datos y Estimaciones de Tipo Estándar

El conjunto de datos elegido para las pruebas corresponde al suelo de La Sal del Rey, región lacustre de baja profundidad, ubicada en el Estado de Texas, EEUU, cuyas coordenadas son (26° 31' 55'' N, 98° 03' 50'' O). Para [16] se obtuvieron ocho muestras integradas por secuencias de ADN correspondientes al gen 16S rRNA. Se encuentran almacenadas en NCBI Short Read Archive bajo el número de acceso SRX008158 [17] La denominación de cada muestra y la cantidad de secuencias que la integra se da en la Tabla 3.1 mientras que el número total de bases químicas almacenadas es de 925673.

Tabla 3.1

Muestra	S85	S86	S87	S88	S89	S90	S91	S91
Tamaño	1641	8361	6926	6146	6226	8444	6103	5885

Con el software libre MOTHUR se realizaron las estimaciones habituales de riqueza y diversidad. Se establecieron las cantidades de especies observadas por suma total de los agrupamientos en OTUs con disimilaridad del 5%. También se calcularon las estimaciones de riqueza CHAO y ACE citadas en 3.2. Con una función de cálculo desarrollada en lenguaje R se evaluaron las respectivas entropías de las muestras. Los resultados se exponen en la Tabla 3.2

Tabla 3.2

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
Cantidad de Secuencias	1641	8361	6926	6146	6226	8444	6103	5885
Entropía	5.938	7.759	7.049	6.852	6.767	6.669	6.687	6.878
Sobs	541	3575	2273	2030	1715	2040	1659	1842
CHAO	834	6351	4080	3363	2502	3180	2691	3110
ACE	1102	9525	5964	3755	2755	3544	3757	4048

4 Simulación Turing

Se utilizaron dos conceptos. Por un lado con la fórmula (2.9) se tiene una estimación de la probabilidad de hallar una especie nueva cuando se selecciona un nuevo individuo para integrar la muestra [9]. Y, por otro, la curva de rarefacción debe alcanzar un comportamiento asintótico horizontal para un tamaño de la muestra suficientemente grande. Esto se produce porque, en realidad, todas las curvas de acumulación a partir de las cuales se construye, observan en más o en menos un

comportamiento asintótico similar cuando aumenta el tamaño muestral. Sobre esta base conjunta se desarrolló el modelo experimental.

Cuando se trata de estimar la cantidad de especies presentes en un medio, 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 el proceso aleatorio que se describe en [3]

Para estimar la probabilidad de que el i-ésimo individuo agregado corresponda a una especie nueva se toma el estimador de Turing

$$\hat{T}_i = \frac{n^{\circ} sgletones}{i-1} \tag{4.1}$$

Entonces, de acuerdo a lo analizado en [3],

$$E(S_i) = S_{i-1}(1-\hat{T}_i) + (S_{i-1}+1)\hat{T}_i \tag{4.2}$$

y operando se obtiene

$$E(S_i) = S_{i-1} + \hat{T}_i \tag{4.3}$$

Si se realiza una simulación eligiendo de a uno individuos en una muestra, cabría esperar que cuando i crezca $\hat{T}_i \rightarrow 0$ pues el número de especies aún no encontradas debiera ir disminuyendo al ser finita la cantidad S de especies buscada. De tal forma también ocurriría que $E(S_i) \rightarrow S$ al crecer el tamaño i.

La simulación se realiza por la técnica de Monte Carlo. Dado el tamaño n de la muestra original se determina el valor del estimador de la probabilidad de especie nueva. Ese valor permite constituir los intervalos $[0, \hat{T}_n)$ y $[\hat{T}_0, 1]$ de modo que al elegir un número aleatorio r tal que $0 \leq r \leq 1$, si cae dentro del primer intervalo el nuevo individuo simulado corresponda a una especie nueva y si cae dentro del segundo intervalo sea un ejemplar de una especie conocida. Si ocurre lo primero, la cantidad de especies en el medio se incrementa en 1 y si no, se utilizan las proporciones existentes de cada especie para asignar por medio de un nuevo número aleatorio la especie ya conocida a la cual pertenece el nuevo individuo. Así se van agregando individuos hasta que la cuenta de las especies nuevas alcance un valor estable. Los pasos del procedimiento se sintetizan en forma secuenciada a continuación.

1- Dada la muestra elegida, de tamaño n, y su agrupamiento en OTUs, se determina el valor inicial del estimador de Turing $\hat{T}_{i+1} = \frac{f_1}{i}$ siendo $i = n$

2- Se elige un número aleatorio r, tal que $0 \leq r \leq 1$ y se pregunta si está en el intervalo $[0, \hat{T}_{i+1})$. Si es así, se realiza $S_{i+1} = S_i + 1$ y se va al paso 4. Si ocurre lo contrario se realiza $S_{i+1} = S_i$ y se va al paso 3

3- Se utiliza la distribución de abundancia de la muestra (ver [3]) para calcular la proporción de individuos que están en OTUs de $1,2,...n$ individuos y con estas proporciones se determina, por un sorteo de

acuerdo a ellas, a que grupo de OTUs ya conocidas pertenece el nuevo individuo. Para establecer a que OTU específica, de entre las de este grupo, corresponde el nuevo individuo se realiza un nuevo sorteo con probabilidad uniforme para cada OTU del grupo.

4- Sea el nuevo individuo de una nueva especie o no, la muestra tiene ahora un elemento más. Se pregunta entonces si el procedimiento debe cortarse porque se cumple el criterio elegido para ello en cuyo caso la simulación ha finalizado. Si el criterio de corte no se cumple, se asigna entonces $i \leftarrow i + 1$, se calcula la nueva distribución de abundancia y la nueva estimación de Turing según $\hat{T}_{i+1} = \frac{f_1}{i}$ y se repite desde el paso 2.

El programa de computadora correspondiente fue desarrollado en lenguaje R [18]. Para fijar el corte del procedimiento se decidió simular casos hasta que la proporción de OTUs singletones, en el total de individuos en la muestra simulada, estuviera por debajo de un cierto umbral. Si esta proporción es baja, ello implica que quedan pocas especies por descubrir y cada vez más agrupamientos u OTUs tienen más de un elemento, conforme avanza la simulación. La simulación se realizó un número fijo de veces para establecer un intervalo de confianza de la estimación de riqueza. Se calculó además la entropía "de salida" o "de corte". Se trabajó sobre el conjunto de muestras SRX008158 con resultados que se expresan en la Tabla 4.1

Tabla 4.1

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
N° Indiv.	1641	8361	6926	6146	6226	8444	6103	5885
Entropía Inicial	5.938	7.759	7.049	6.852	6.767	6.669	6.687	6.878
Sobs	541	3575	2273	2030	1715	2040	1659	1842
CHAO	834	6351	4080	3363	2502	3180	2691	3110
ACE	1102	9525	5964	3755	2755	3544	3757	4048
N° Simul.	7	7	7	7	7	7	7	7
Confianza	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Media Simulada Corte=0.03	1288	9851	6015	5567	3759	4384	3924	4800
Intervalo	1215-1361	9542-10159	5706-6323	5381-5753	3646-3872	4237-4532	3816-4032	4542-5058
Individuos Simulados Promedio	10491	60000	51615	48292	29859	35728	32989	41156
Entropía Final	6.304	8.326	7.546	7.387	7.106	6.995	7.071	7.352

El análisis de los valores detallados permite establecer que la estimación promedio de las simulaciones Turing, utilizando una proporción de corte de 0.03, supera en todos los casos a la estimación respectiva realizada por ACE.

Además salvo para la muestra S87, para todas las otras, la estimación ACE está por debajo del límite inferior del intervalo de confianza del 95% construido para la estimación de la riqueza de especies según este método. Para todas las muestras, la estimación de CHAO queda sensiblemente por debajo del límite inferior del intervalo de confianza del 95% establecido para la estimación del número de especies, efectuada de acuerdo a este procedimiento. En el caso de la muestra S86, las cantidades promedio de individuos revelan que no fue alcanzado el umbral de corte previsto a pesar de lo cual las estimaciones estuvieron a tono con las realizadas para las otras muestras. Por último, la entropía final promedio, calculada sobre las muestras, destaca un crecimiento respecto de la exhibida por las muestras reales iniciales. Este incremento revela el aumento de la diversidad producido por el agregado de especies.

5 Simulación Entropía

De la Tabla 4.1 se extraen dos conclusiones importantes. La primera registra que la entropía crece según aumenta la cantidad de individuos simulados y la segunda constata que la diferencia entre los valores de entropía de dos pasos sucesivos de la simulación va disminuyendo conforme aumentan los casos simulados y se va convergiendo a los valores de riqueza poblacionales. La entropía es la medida de diversidad más comúnmente usada [19] y los efectos apuntados se utilizaron para construir el coeficiente

$$B_i = 1 + H_i \left| H_i - H_{i-1} \right| \tag{5.1}$$

que corrige la probabilidad de especie nueva

$$\hat{T}_i corr = B_i \hat{T}_i \tag{5.2}$$

Así se tiene en cuenta que el valor de la entropía crece con la cantidad de individuos simulados que disminuyen las diferencias entre las entropías del paso actual y del anterior conforme avanza la simulación. La entropía se calcula

$$H_i = -\sum_{j=1}^{S_i} \hat{p}_j \log \hat{p}_j \tag{5.3}$$

con $\hat{p}_j = \frac{i_j}{i}$. La secuencia de programación en lenguaje R continúa como fue detallado en 4. Las pruebas sobre el conjunto SRX008158 condujeron a los valores de la Tabla 5.1

Tabla 5.1

Muestra	S85	S86	S87	S88	S89	S90	S91	S92
N° Individ.	1641	8361	6926	6146	6226	8444	6103	5885
Entropía Inicial	5.938	7.759	7.049	6.852	6.767	6.669	6.687	6.878
Sobs	541	3575	2273	2030	1715	2040	1659	1842
CHAO	834	6351	4080	3363	2502	3180	2691	3110
ACE	1102	9525	5964	3755	2755	3544	3757	4048

N° Simul.	7	7	7	7	7	7	7	7
Confianza	0.95	0.95	0.95	0.95	0.95	0.95	0.95	0.95
Media Turing Tabla 4.1	1288	9851	6015	5567	3759	4384	3924	4800
Media con Entropía/ Corte=0.03	1294	9880	5997	5393	3735	4401	3860	4781
Intervalo	1178-1410	9711-10049	5797-6198	5172-5615	3681-3790	4310-4492	3740-3980	4609-4952
Individuos Simulados Promedio	10500	60000	51269	45874	29856	36381	32127	41074
Entropía Final	6.296	8.329	7.538	7.364	7.104	7.006	7.054	7.341

Los valores estimados por la simulación con entropía son también mayores que los obtenidos por la estimación ACE y guardan cercana relación con los calculados por la simulación sin corrección por entropía. Esta corrección involucra en cada muestra solo unas decenas de especies lo que puede interpretarse como un ajuste "fino" de la estimación. En el caso de la muestra S86 la cantidad promedio de individuos es 60000 pues ese es el límite que se impuso en el programa a los individuos simulados y, al alcanzarlo, el valor de corte fue de 0.043.

6 Conclusiones

En primer lugar se destaca que el método de simulación aplicado a la estimación de la riqueza ha mostrado su capacidad para la tarea, al producir resultados coherentes, que están dentro del orden de los hallados por la estimación no paramétrica y por rarefacción, pero lo suficientemente mayores como para disminuir la subestimación que esos métodos producen. La comparación de la simulación Turing con ACE ha sido siempre ventajosa, bajo la condición de seleccionar un adecuado valor umbral para realizar el corte de la misma. Esta diferencia entre ambas estimaciones resulta útil para señalar que hay una pérdida de diversidad al realizar la estimación ACE que tiene en cuenta la cantidad esperada de singletones f_1 solo en la muestra real inicial. La estimación por simulación Turing va reduciendo de estado a estado esta cantidad, de forma que la muestra simulada correspondiente a la iteración final produce una mejora en la evaluación del número de especies presentes en el medio. En segundo término cabe señalar que al considerar la variación de la diversidad a través de la entropía, utilizándola de la manera descripta para corregir la estimación de la probabilidad de nueva especie en cada iteración, se obtienen valores de riqueza similares, o muy levemente superiores que representan un ajuste de la estimación a las condiciones de uniformidad-dominancia de especies halladas en la muestra.

Bibliografía

1. 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. (2008)
2. Hughes, J, Hellmann, J, Ricketts, T y Bohannan, B. "Counting the uncountable: statistical approaches to estimating microbial diversity". Appied and Environmental Microbiology. 4399-4406. (2001)
3. Santa María, C. y Soria, M. "Aplicaciones de Data Mining al Estudio de la Biodiversidad". WICC2011. BDMD. 3773. (2011)
4. Magurran, A. Measuring Biological Diversity. Blackwell Science Ltd. (2004)
5. Bunge, J. y Fitzpatrick, M. "Estimating the Number of Species: A Review". Journal of American Statistical Association. Vol. 88. N° 421. pp. 364-373. (1993)
6. 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. (2003)
7. Chao, A. "Nonparametric estimation of the number of classes in a population". Scand J Statist 11: 265-270. (1984)
8. Chao, A y Lee, S. "Estimating the Number of Classes via Sample Coverage". Journal of American Statistical Association. Volume 87. Issue 417. (1992)
9. Good, I. "The Population Frequencies of Species and Estimation of Population Parameters". Biometrika. Vol 40 N° 3/4. (1953)
10. Hughes,J y Hellman, J. "The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity". Methods in Enzymology. Vol 397. (2005)
11. Efron, B. "Computers and theory of statistics: thinking the unthinkable". Technical Report N° 39. Division of Biostatistics. Stanford University (1978)
12. Hughes,J y Hellman, J. "The Application of Rarefaction Techniques to Molecular Inventories of Microbial Diversity". Methods in Enzymology. Vol 397. (2005)
13. Tellinghuisen, J. "The Least_Squares Analysis of Data from Binding and Enzyme Kinetics Studies: Weights, Bias, and Confidence intervals in Usual and Unusual Situations". Methods in Enzymology. Volume 467. Pgs.500-527. (2009)
14. O'Hara, R. "Species richness estimators: how many species can dance on de head of a pin. Journal of Animal Ecology. 74, 375.386. (2005)
15. 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. (2007)
16. 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. (2010)
17. http://www.ncbi.nlm.nih.gov/
18. http://www.r-project.org/
19. Hill, T, Walsh, K, Harris, J yMoffett, B. "Using Ecological Diversity Measures with Bacterials Communities". FEMS.Microbiology Ecology 43 1-11 (2003)