MINERÍA DE DATOS SOBRE COMUNIDADES BIOLÓGICAS

Cristóbal R. Santa María Departamento de Ingeniería UNLAM
Marcelo Soria Facultad de Agronomía Cátedra de Microbiología UBA

RESUMEN

La práctica científica y tecnológica suele reunir conceptos originados en diversas disciplinas para desarrollar perfiles y potenciales usos que adquieren cierta unidad e independencia conceptual. Tal es el caso de data mining que a partir de la tecnología de las bases de datos incorporó paulatinamente ideas provenientes de la inteligencia artificial y de la estadística para clasificar y/o predecir resultados sobre un muy variado conjunto de sistemas. El proyecto de investigación aquí presentado estudia técnicas bioinformáticas con las que se trabaja sobre comunidades microbiológicas de suelos. Tales métodos tienen el propósito de clasificar los organismos que forman parte del medio y predecir su diversidad. El análisis parte de la representación computacional del ADN que codifica la información genética y establece, con datos obtenidos a partir de muestras, las propiedades del conjunto de microorganismos que conforman esa comunidad. Este tipo de estudio, denominado metagenómica, permite agrupar los distintos tipos de organismos en clusters que representan alguna categoría taxonómica como especie, género, familia etc. También es posible a partir de estos agrupamientos realizar estimaciones de biodiversidad que proporcionen información sobre la potencialidad y riqueza del suelo.

El proyecto de investigación tiene dos objetivos. Por un lado establecer un modelo bioinformático markoviano para la comparación de secuencias de ADN a efecto de clasificación, y por otro presentar un análisis crítico de los procedimientos de data mining aplicados a la evaluación de la riqueza en distintos ecosistemas.
Palabras Clave: metagenómica cluster predicción adn modelo markoviano biodiversidad

CONTEXTO

El trabajo se inscribe dentro de la línea de investigación Modelos Bioinformáticos de Markov para Vías Metabólicas en Metagenomas que se lleva adelante en la UNLAM dentro del Programa de Incentivos y con la orientación brindada desde la Maestría en Explotación de Datos y Descubrimiento del Conocimiento de la Facultad de Ciencias Exactas y Naturales de la UBA.

1. INTRODUCCION

ADN

En forma esquemática puede decirse que la biología molecular investiga la estructura y función de las proteínas y los ácidos nucleicos en los organismos vivos. Esto incluye necesariamente el estudio de los genes que integran el genoma de cada organismo y que se encuentran presentes en los cromosomas dentro de las células. El ADN contiene la información necesaria para que la célula sintetice las proteínas que, a su vez, dan forma a los organismos o funcionan como catalizadores de reacciones metabólicas.

La estructura del ADN es una doble cadena helicoidal formada por las bases químicas enfrentadas A (adenina) – T (timina) y C (citosina) – G (guanina), de modo tal que una cadena resulta químicamente complementaria de la otra. Por lo tanto,

para la búsqueda genética puede utilizarse una sola cadena al sobreentender la presencia de la complementaria. Las llamadas técnicas de secuenciación transforman la estructura química en información computacional constituida por secuencias de letras que representan las bases químicas detalladas. Así una parte de una secuencia ejemplo puede ser: …ATTGGTACCGAT…La cantidad de bases o nucleótidos que contiene una secuencia completa depende del organismo, pero el número suele ser del orden de los cientos de millones. Los genes se disponen sobre una o más moléculas grandes de ADN conocidas como cromosomas. Cada gen ocupa un lugar específico en el cromosoma llamado "locus" (locus, loci: lugar en latín). Las cadenas de ADN cromosómico también contienen pares de bases que no son constitutivos de genes y que pueden señalar zonas de separación o tener otras funciones, incluso aún desconocidas. Además la parte de la secuencia que caracteriza a un gen puede presentar segmentos que no se utilizan estrictamente para codificar proteínas o contener, en organismos distintos, inserciones, ausencias o reemplazos de bases. Tales modificaciones a veces no alteran la codificación de la proteína correspondiente pero, en otros casos, pueden indicar mutaciones genéticas que cambian la codificación.

Las proteínas son a su vez cadenas de aminoácidos de largo variable. Existen 20 aminoácidos que constituyen las proteínas. A nivel del ADN, cada aminoácido está codificado por sucesiones de tres bases llamadas codones. Una rápida cuenta indicaría que en principio hay 4^3=64 codones que codifican para los 20 aminoácidos diferentes considerando todas las posibles ternas integradas por las cuatro letras A,T,C,G. Sin embargo, esto no es así. Cada uno de los 20 aminoçidos está representado por uno o varias de estas ternas o codones. A su vez, hay ternas que no representan aminoácidos sino señales de parada o finalización. El llamado código genético muestra la correspondencia entre codones y aminoácidos. Para facilitar la representación informática cada aminoácido puede ser nombrado por una letra de forma tal que la cadena original de bases del ADN puede reemplazarse por una secuencia de letras que representan los aminoácidos [1].

Las secuencias de ADN obtenidas en cualquier estudio pueden ser comparadas con otras cuya forma y función ya ha sido "anotada" en una base de datos. Existen a nivel internacional varias bases de datos de carácter público, por ejemplo la GenBank del National Center for Biotechnology Information (NCBI), que permiten la consulta on-line por Internet y que contienen tanto secuencias de bases como de aminoácidos. Al realizar la comparación se busca hallar regiones que respondan a segmentos de cadena conocidos. Este mecanismo de búsqueda de patrones se denomina "alinear" la secuencia y puede realizarse para una sola cadena o para varias simultáneamente. El algoritmo mas utilizado en esta tarea es el BLAST (Basic Local Alignment Search Tool **)** que es de tipo heurístico y asigna un puntaje al alineamiento producido [2]. Este puntaje indica el grado de similitud de la cadena analizada con alguna otra, cuya función o presencia en un tipo de organismo ya ha sido identificada y anotada en la base de datos. En forma más o menos reciente se han desarrollado métodos de alineamiento sobre la base de modelos ocultos de Markov que se entrenan para reconocer distintos tipos de dominios e incluso permiten la aplicación conjunta con BLAST. Con estas técnicas se trabaja para elaborar el mapa genético de una

especie que establece el "locus" de cada gen en el cromosoma y en que par homólogo de cromosomas se encuentra, siendo que los cromosomas se presentan de a pares y que las diferentes especies pueden presentar, además, cantidades distintas de pares de cromosomas. Esta información contribuye para determinar la función de un gen y las características de "proximidad" e independencia con otros, lo que a su vez influye en la recombinación genética producida por la división celular (mitosis).

METAGENÓMICA

La metagenómica, cuyo desarrollo comienza con el actual siglo, realiza el análisis genómico de comunidades microbianas [3]. Combina el concepto estadístico de meta-análisis referido al proceso en el que se relacionan estadísticamente análisis separados, con la genómica que es el análisis comprensivo del material genético de un organismo. Este nuevo campo trata de explorar un conjunto de datos constituido por fragmentos de ADN de tamaño variable originados en genomas de distintos organismos. Estos organismos deben realizar un gran número de actividades metabólicas para sobrevivir y multiplicarse en sistemas o ambientes tales como el digestivo humano, el medio acuático marino o el suelo fértil. Las enzimas responsables de estas actividades metabólicas están codificadas en el ADN genómico y por tanto se pueden recuperar mediante el análisis metagenómico.

Si se tiene en cuenta que el número total estimado de procariotas (organismos unicelulares sin núcleo) presentes en el planeta es mayor que 10^{29} se comprende que, en los distintos ecosistemas biológicos, se hallen en cantidades considerables [4]. Todos estos procariotas poseen una estructura genética y algunos tienen especial incidencia en los procesos de transformación química que ocurren en el medio que habitan. En muchos casos, no es posible cultivarlos en laboratorio a efecto de extraer su ADN e investigar su genoma en forma aislada. El conocimiento microbiológico obtenido por técnicas de laboratorio que no incluyen la secuenciación y el alineamiento computacional de ADN alcanza sólo al 1% de los microorganismos presentes en un ecosistema [5]. Sin embargo, al tomar muestras heterogéneas del medio, pueden secuenciarse cadenas de ADN que contienen fragmentos de los distintos microorganismos presentes y comparar estos fragmentos con secuencias genéticas anotadas en distintas bases de datos. Así se van identificando genes y conjuntos de genes con el propósito de establecer el genoma y la función de una proporción mayor de entidades microbianas.

En muchos casos la distinción entre especies es claramente señalada por disimilaridades aparecidas en secuencias correspondientes a un gen en particular. Estas secuencias se elijen porque, en general, han sido bien conservadas en el desarrollo evolutivo y presentan mínimas variaciones que indican la diferencia de especies. Tal es el caso del gen 16S rRNA, el más común de estos marcadores [3], presente en procariotas. Este gen es utilizado en estudios llamados filogenéticos que tratan de la evolución y del desarrollo de las especies. Se comparan secuencias y se ven las diferencias para estructurar árboles filogenéticos y secuencias evolutivas al partir de una división en ramas denominadas eucariotas, bacterias y arqueobacterias, estas dos últimas procariotas.

Un aspecto de interés al considerar la metagenómica, es la estimación de la riqueza de la comunidad analizada en términos de

cantidad de especies presentes y caminos metabólicos que estas recorren.

SECUENCIACION Y CONTIGS

Obtener una secuencia de letras que represente la cadena de bases químicas que integran una molécula de ADN requiere una tecnología que se encuentra actualmente en permanente cambio y que, merced a ello, logra establecer secuencias cada vez mas largas y mas precisas con mayor rapidez. El desarrollo de estas técnicas constituye uno de los pilares que fundamentan los avances en la determinación completa de genomas y, por ende, en el conocimiento de la función y el comportamiento de diversos organismos en un ecosistema. Como consecuencia, ello ha dado lugar a muy distintas aplicaciones en campos tales como la prevención y detección de enfermedades, la fabricación de medicamentos, el mejoramiento de razas animales o de especies vegetales que se utilizan para la alimentación humana, la creación de bancos genéticos para la identificación de personas y otras. Las cadenas correspondientes a los genes de variados organismos se anotan en grandes bases de datos donde pueden consultarse para diversos fines.

Esta reunión de la biología molecular con la computación ha dado lugar a una nueva disciplina denominada bioinformática para la cual se han desarrollado modelos matemáticos y estadísticos originales y procedimientos de cálculo surgidos de las necesidades de determinación biológica [2].

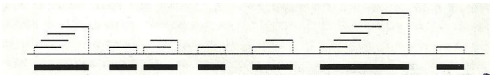
La primera técnica de secuenciado para transformar la molécula de ADN en una sucesión de símbolos, estuvo disponible en 1975 desarrollada por Frederic Sanger et al y se la conoce como método de terminación de cadena. Basada en separación electroforética, su costo es aproximadamente de US $1 por cada 1000 pares de bases (en adelante se abrevia bp), su velocidad es de alrededor de 24 bases por segundo y puede alcanzar una precisión en el secuenciado de 99.9%. A partir de 1996 se comenzó a aplicar la técnica de pirosecuenciación conocida también por su nombre comercial 454. Está fundada en el uso de fibras ópticas que transmiten quimioluminiscencia. Permite secuenciar alrededor de 11000 bases por segundo a un costo de US $ 0.035 por cada 1000 bp. De precisión similar al método Sanger agrega las ventajas comparativas de ser mucho más veloz y sensiblemente más barata. A principios del año 2009 se contabilizaban, en el orden mundial, más de 130 proyectos metagenómicos que se ejecutaban utilizando pirosecuenciación de ADN [5].

A diferencia de los estudios genómicos, en los que se parte de muestras de ADN de un solo tipo de organismo, en la metagenómica las muestras contienen ADN de los distintos tipos de organismos presentes en el medio. Por técnicas de laboratorio se aísla el ADN mezclado y se lo fragmenta insertándolo en vectores. Estos insertos constituyen el material que es secuenciado. Las secuencias obtenidas contienen entonces fragmentos de ADN correspondientes a distintos organismos.

En un proyecto metagenómico el flujo del análisis de datos comienza con el ensamblado de tales secuencias a efecto de obtener cadenas de mayor longitud y ganar así profundidad en la estructura de la población [6]. Secuencias cortas, de alrededor de 500 nucleótidos, obtenidas por pirosecuenciación, son superpuestas para elaborar una sola denominada *contig* [7]. Esta metodología llamada *shotgun*, aplicada sobre secuencias extraídas de muestras metagenómicas,

reconstruye segmentos más largos de ADN correspondientes a diferentes especies del ecosistema. El procedimiento afronta la dificultad de que no es conocida la localización de los fragmentos en la supuesta cadena que integra el ADN de todos los organismos pero, aún así, al superponer suficientes fragmentos es posible armar una secuencia larga donde varios contigs cubren una proporción de la cadena teórica. Si la longitud de los fragmentos es L y la de la secuencia obtenida finalmente resulta un múltiplo nL se dice que se obtuvo *cobertura* n. La Figura 1 ilustra la idea al mostrar en línea gruesa los distintos contigs obtenidos a partir de diferentes cantidades de fragmentos.

Figura 1



2. LINEAS DE INVESTIGACION y DESARROLLO

Se trabaja actualmente sobre dos temas.

El primero se refiere al diseño, entrenamiento y testeo de un modelo oculto de Markov capaz de detectar la codificación de enzimas que intervienen en una ruta del metabolismo metagenómico en suelos agrarios. Esto puede tener varios usos:

i) Conocer con que probabilidad una secuencia genómica de un microorganismo del ecosistema codifica para esas enzimas.

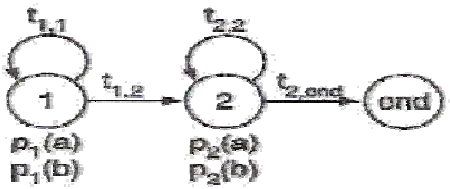
ii) Conocer con que probabilidad las enzimas se hallan presentes en las distintas reacciones de una vía metabólica cuando se conocen los genomas de los microorganismos intervinientes.

iii) Ayudar a conocer y determinar el genoma de microorganismos que se sabe están presentes en las reacciones metabólicas del ecosistema.

El esquema de modelado desarrollado por Leonard Baum y aplicado inicialmente al reconocimiento del habla, fue utilizado posteriormente con éxito en bioinformática. Se trata del Modelo Oculto de Markov. Este modelo, nombrado en la bibliografía con las siglas HMM, supone una variable aleatoria para cada elemento de una secuencia de bases o aminoácidos. Se tiene entonces una sucesión de estados cuyos valores observables son las bases químicas A,T,C,G o alternativamente las letras que identifican a los 20 tipos de aminoácidos. El supuesto markoviano estriba en que el estado actual (el enésimo de la sucesión de estados) solo depende del anterior y se alcanza de acuerdo a una determinada probabilidad de transición entre estados. La cadena de estados que representa la secuencia de un gen (u otro tipo de secuencia buscada) está en realidad oculta para el analista que solo tiene a la vista la secuencia que desea "alinear" obtenida en el laboratorio. En esta secuencia visible, cada letra correspondiente a una base química o a un aminoácido se manifestará en ese estado con una probabilidad denominada de emisión. Tal probabilidad variará además de acuerdo a lo que el modelo elegido postula para cada estado.

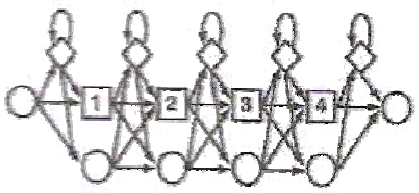
El número de estados considerados, sus probabilidades de transición entre estados y de emisión en cada estado constituyen el modelo según se ilustra en la Figura 2.

Figura 2



Se puede apreciar que el modelo consta de dos aspectos diferentes: la arquitectura y los parámetros. La arquitectura tiene en cuenta cuales y cuantos son los estados que interesa modelar. En la Figura 3 los cuadrados indican los estados que deben considerarse presentes, los rombos aquellos estados que eventualmente puedan alcanzarse por inserciones en la cadena de símbolos y los círculos corresponden a ausencias u omisiones que puedan presentarse en la cadena o al comienzo y fin de la misma [8]

Figura 3



El diseño de un esquema tal constituye el primer paso en la construcción del modelo y depende de los conocimientos o de las suposiciones que se hagan "a priori" sobre las cadenas y la información que codifican. Es una tarea de elaboración teórica.

Luego hay que determinar las distintas probabilidades de transición entre estados y las probabilidades de emisión en cada uno de ellos, las que constituyen el juego de parámetros del modelo. Este trabajo es empírico pues se apoya en inferencias estadísticas sobre muestras de cadenas y se realiza por entrenamiento y testeo en un proceso denominado de aprendizaje automático.

Es claro que una vez cumplidas las etapas de diseño y de selección de parámetros pueden plantearse diseños alternativos y hallarse sus respectivos juegos de parámetros de modo de comparar el desempeño de los distintos modelos construidos.

Se trabaja entonces con la idea de desarrollar un "perfil" HMM que represente la cadena de estados de interés aportando mayor precisión en la identificación y/o la comparación de estructuras genómica presentes en suelos dedicados al trabajo agrario [9]. Tal condición obligará a la mezcla jerárquica de cadenas de Markov para elaborar un único modelo probabilístico que tenga en cuenta las variaciones de forma que se presentan en la codificación de las proteínas en distintos organismos [10] , [11].

El segundo tema abordado se refiere al clustering de microorganismos según categorías taxonómicas y al análisis del tipo de predicción que se realiza sobre la biodiversidad de los suelos señalados.

Ya se ha mencionado que el gen 16S rRNA se utiliza como marcador pues se ha conservado a través de la cadena evolutiva. Se presenta en bacterias y arqueobacterias y la secuencia del gen sufre cambios que responden a diferencias biológicas entre organismos. Si se establece una forma para medir la similaridad entre dos secuencias, se pueden agrupar los tipos de microorganismos según un porcentaje que representa su grado de parecido. De esta forma se asignan a un mismo grupo todas las secuencias cuyas diferencias no exceden un porcentaje prefijado.

El primer paso para cumplir esta tarea es definir una "distancia genética" que permita representar la diferencia entre dos secuencias. A continuación se construye una matriz de distancias utilizando todas las secuencias disponibles. Cada celda de esta matriz es la distancia que hay entre la secuencia de la fila y la de la columna. A tal efecto existen programas aplicables on-line que calculan esta matriz utilizando distintas métricas. Por

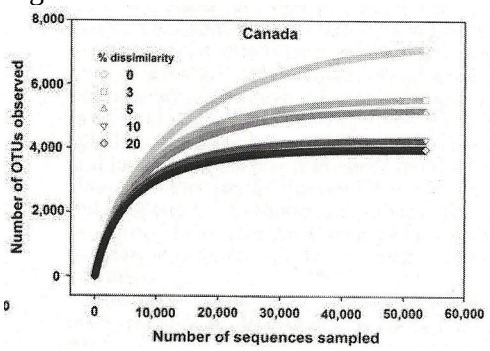
ejemplo, Schloss y Handelsman utilizaron el software ClustalW para alinear las secuencias y construir la matriz de distancias con el programa DNADIST del paquete PHYLIP [12].

A continuación se fija el criterio con el que dos secuencias se considerarán similares. En este punto es importante aclarar que es habitual usar el porcentaje de disimilaridad complementario que resalta la diferencia y no el parecido. En cualquier caso hay que tener en cuenta que la taxonomía reconoce como jerarquías de orden creciente a especie, género, familia, orden, clase, phylum y dominio. Los distintos porcentajes de disimilaridad indican la máxima diferencia que puede existir entre cadenas de 16S rRNA correspondientes a los individuos de un grupo. Si bien en términos biológicos no puede aplicarse estrictamente el criterio computacional enunciado, 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 para otra menor que el 20% hay igual clase o phylum. Los grupos así obtenidos se denominan Unidades Taxonómicas Operacionales que se citan subindicando el porcentaje referido. Por ejemplo $OTU_{3\%}$ u $OTU_{20\%}$

Para armar estas OTUs se pueden emplear formas de agrupamiento tales como vecinos más cercanos o vecinos del promedio que responden a distintas formas de enfocar las diferencias. Esto también puede realizarse por medio de programas que se bajan de la web tales como MOTHUR elaborado por Schloss y cols. [12]

La biodiversidad presente en un ecosistema se suele medir por medio de la llamada riqueza, que nos es otra cosa que el número total de especies presentes en el medio. Si se tiene en cuenta la enorme cantidad de microorganismos que pueden estar presentes en un suelo cultivado se comprende de inmediato lo difícil que resulta conocer ese número con cierta precisión [13]. El problema es que sencillamente no se tiene "a priori" una estimación del número de especies y en muchos casos tampoco se cuenta con su orden de magnitud aproximado [14]. Esto trae como consecuencia que no se pueda asegurar que el tamaño de las muestras que se utilizan en la estimación sea el adecuado a efecto de evaluar la riqueza del medio. Se utilizan entonces curvas de rarefacción basadas en procedimientos de remuestreo tales como bootstrap y jackniffe que toman en cuenta los promedios de OTUs obtenidos según diferentes tamaños de las muestras. A continuación se evalúa el tamaño de muestras para el cual la curva exhibe un comportamiento asintótico y se adopta tal tamaño como el adecuado [15]. Este concepto se ilustra en la Figura 4.

Figura 4



En realidad las curvas proceden de ajustes estadísticos efectuados sobre los datos, que se realizan con distribuciones tales como Lognormal, Pareto, Gamma o Gaussiana inversa. [16]. Los ajustes se efectúan por medio del test Chi-Cuadrado que requiere ciertos cuidados técnicos no siempre observados [17].

Por otra parte se presentan también inconvenientes en la suposición inicial sobre la distribución de especies en el medio. La hipótesis de

comportamiento uniforme de las mismas está lejos de ser realista pues en todo ecosistema hay especies dominantes en número. Se verifican en suma dificultades de orden biológico y de procedimiento que aparecen al tratar de cuantificar la biodiversidad [13]. Estas dificultades sugieren la investigación de tres aspectos de la estimación:

i) Estudio de hipótesis adecuadas de distribución de especies.
ii) Evaluación de metodologías para establecer el tamaño muestral a partir de curvas asintóticas de rarefacción.
iii) Análisis y refinamiento de los procedimientos de ajuste estadístico utilizados.

3. RESULTADOS OBTENIDOS/ESPERADOS

El resultado obtenido hasta ahora se relaciona con la lectura y discusión de la bibliografía que se cita en el apartado 5 y con la identificación de los puntos de interés para la investigación dentro de la sucinta descripción realizada. Cabe aclarar que el aspecto abordado es bastante nuevo y que en nuestro medio recién comienzan a verse los primeros trabajos sobre el tema.

En cuanto a los resultados esperados para esta etapa de trabajo se cita especialmente la elaboración de un análisis crítico detallado de los procedimientos de evaluación de la riqueza de ecosistemas con propuestas que mejoren la precisión y efectividad de las estimaciones y la aplicación a muestras tomadas en suelos locales. También se espera comenzar un estudio comparativo de diferentes Modelos de Markov desarrollados desde distintas bases de datos internacionales a efecto de acumular conocimiento para la formulación de un modelo ajustado a los datos locales.

4. 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. Colabora además en aspectos estadísticos del trabajo la Especialista María Eugenia Ángel, quien obtuvo igual especialización en 2008 y es docente e investigadora en la UNLAM.

5. BIBLIOGRAFIA

[1] [1997] Setubal, J y Meidanis, J. Introduction to Computational Molecular Biology. PWS Publishing Company.

[2] [1998] Durbin, R, Eddy, S, Krogh, A y Mitchison, G. Biological Sequence Analysis. Cambridge University Press.

[3] [2007] Gross, L. "Untapped Bounty: Sampling the Seas to Survey Microbial Biodiversity". PLoS Biology/ Volume 5/Issue 3/e85

[4] [2006] Grasso, D. "Metagenómica: un viaje a las estrellas". Revista Argentina de Microbiología. Volumen 38. N° 4.

[5] [2009] Guazzaroni, M.E, Beloqui, A, Golyshin, P y Ferrer, M. "Metagenomics as a new technological tool to gain scientific knowledge".

World Journal Microbiology Biotechnology. 25:945-954
[6] [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
[7] [2005] Ewens, W y Grant,G. Statistical Methods in Bioinformatics: An Introduction. Springer Science+Business Media, Inc.
[8]- [1998] Eddy, Sean R. "Profile hidden Markov models". Bioinformatics Review. Vol. 14 n° 9. Pgs. 755-763.
[9]- [2006] Winters-Hilt, Stephen. "Hidden Markov Model Variants and their Application". BMC Bioinformatics. 7(Suppl 2):514.
[10]- [2004] Hiroshi Mamitsuka y Yasushi Okuno. "A Hierarchical Mixture of Markov Models for Finding Biologically Active Metabolic Paths using Gene Expression and Protein Classes". Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference (CSB) 0-7695-2194-0/04.
[11]- [2004] Zhang, Nevin L. "Hierarchical Latent Class Models for Cluster Analysis". Journal of Machine Learning Research 5 697-723
[12]- [2009] Schloss, Patrick D., Westcott, Sarah L., Ryabin, Thomas, Hall, Justine R., Hartmann, Martin, Hollister, Emily B., Lesniewski, Ryan A., Oakley, Brian B., Parks, Donovan H., Robinson, Courtney J., Sahl, Jason W., Stres, Blaz, Thallinger, Gerhard G., Van Horn, David J., Weber, Carolyn, F. "Introducing mothur: Open-Source, Platform-Independent, Community-Supported Software for Describing and Comparing Microbial Communities" Appl. Environ. Microbiol. 75:7537-7541
[13] [2001] Gotelli, N y Colwell, R. "Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness". Ecology Letters 4:379-391.
[14] [2006] Schols, P y Handelsman, J. "Toward a Census of Bacteria in Soil". PLoS Computational Biology. Vol 2/Issue7/e92.
[15] [2007] Roesch, L, Fulthorpe, R et al. "Pyrodequencing enumerates and contrasts soil microbial diversity". ISME Journal 1, 283-290.
[16] [2006] Hong, S, Bunge, J, Jeon, S, Epstein, S. "Predicting microbial species richness". PNAS. Vol. 103 N°1 Págs. 117-122
[17] [1988] Canavos, G. Probabilidad y Estadística. McGraw-Hill
[18] [2007] Bergeron, A, Belcaid, M, Steward, G y Poisson,G. "Divide and Conquer: Enriching Environmentals Sequencing Data". PLos One 2(9); e830.doi:10.1371/journal.pone.0000830
[19]- [2006] Borodovsky, Mark y Ekisheva, Svetlana. Problems and Solutions in Biological Sequence Analysis. Cambridge University Press.