RELACIÓN TAXONÓMICA DE LAS ESPECIES BACTERIANAS CON LOS MEDIOS ECOLOGICOS MEDIANTE BASES DE DATOS RELACIONALES

MIKEL AGUIRRE RODRIGO

MAGISTER EN BIOINFORMÁTICA Y BIOLOGIA COMPUTACIONAL
UNIVERSIDAD COMPLUTENSE DE MADRID

2011-2012

CNB-CSIC

JAVIER TAMAMES DE LA HUERTA

SEPTIEMBRE DEL 2012

CALIFICACIÓN:

Relación taxonómica de las especies bacterianas con los medios ecológicos mediante bases de datos relacionales

Índice:

Objetivos:

El objetivo final de todo este trabajo es poder analizar en conjunto diferentes interacciones entre especies y poder cuantificar estos datos en base a toda la información que se dispone en la base de datos del Genbank. Con el fin de poder ver todas las posibles interacciones que existen entre diversas especies, es necesario realizar una extracción adecuada de toda la información que se pueda lograr de las entradas del Genbank, puesto que a la hora de determinar ciertas muestras se consideraran varios factores. Por ello ante de llegar al objetivo final del trabajo hay que realizar varias tareas previas para un resultado final satisfactorio:

-Crear una base de datos obteniendo os datos de interés de la base de datos del Genbank.

-Analizar cada dato obtenido en esa base de datos con el fin de obtener mas información de los mismos. Como la existencia de las diferentes especies que se puedan identificar dentro de la base de datos.

-Clasificar taxonómicamente todas las secuencias existentes en el Genbank que hagan referencia al gen 16S.

Una vez identificadas las especies de microorganismos, clasificarlas por medios para después poder analizar las posibles redes de interacción que puede haber entre dichos seres vivos. Este proceso se podrá llevar a cabo después de tener toda la base de datos llena y actualizada. Ya que cuanto mas nuevo datos haya mas fiable será la información que se pueda conseguir de la base de datos.

El programa de relación entre los taxones y las muestras no se ha desarrollado en este proyecto. Pero se han dejado todas las herramientas a disposición del usuario que lo lleve a cabo.

Gracias a esta base de datos relacional podremos sacar información de la evolución de las entradas a lo largo del tiempo. Como por ejemplo la evolución de la entrada de las secuencias que se han introducido en el Genbank en diferentes años. Y de la misma forma cuantas especies diferentes se pueden encontrar en base al 16S y la evolución de estas entradas a lo largo del tiempo.

También hay que tener en cuenta las actualizaciones que puedan ocurrir en las otras beses de datos que se utilizan en el proceso, como la de greengenes. Puesto que el conocimiento de nuevas taxonomías y clasificaciones pueden cambiar la asignación de las secuencias a diferentes taxones.

Introducción:

Hace años que se habla de diferentes redes de interacciones entre diferentes microorganismos. De hecho ya se han descrito diferentes relaciones entre diferentes microorganismos en diversos ecosistemas. En los casos de ecosistemas microbianos mas cerrados observamos unas redes de interacción muy fuertes ,como puede ser el caso del intestino humano[1][2][3]. Este caso es uno de los ecosistemas donde se ha caracterizado de una forma mas completa la estructura microbiana, junto con otras partes del cuerpo humano en las que las comunidades bacterianas están presentes, como la cavidad oral, la piel, las cavidades respiratorias, etc [2][3]. En estos casos las mayoría de los taxones y la relación que existen entre ellos ha sido descrita de forma muy exhaustiva y completa [3]. Pero esto ha sido posible gracias a ser ecosistemas lo suficientemente cerrados como para que las interacciones entre las especies este muy marcada. Este tipo de estructura lleva a que el simple hecho de la entrada de un nuevo microorganismo provoque la modificación de las redes de interacción que existen entre ellos.

Por otro lado, tenemos los casos en el que las relaciones entre los microorganismos no es tan marcada; como es el caso de aguas marinas de la costa [4], ya que las simples corrientes modifican el entorno constantemente. Pese a ello en estos casos, aunque el limite del ecosistema microbiano no está muy bien delimitado y las relaciones entre las especies sea mas débil, se han constatado la existencia de diferentes relaciones entre diversas especies. En este caso, además, se han encontrado nuevos y extraños taxones, lo que demuestra que con las nuevas tecnologías que se están desarrollando se pueden describir nuevas especies [4].

En estos momentos, gracias a la cantidad de datos a la que se tiene disposición podemos analizar los taxones que se relacionan con diversos ecosistemas [5]. Estos trabajos, que se basan en el análisis de grandes cantidades de datos necesitan de una descripción lo mas completa posible para que sean fiables.

En este caso, la razón de querer clasificar de forma adecuada las especies por ecosistemas y funcionalidad reside en entender mejor la base microbiológica de cada ecosistema. Para ello, la necesidad de lograr la mayor información respecto al medio en el que se extraído es primordial. Ya que cuanta mas información se disponga del medio mas acertada será la descripción del mismo y menos probabilidades habrá de equivocarse. Este suele ser un gran problema porque en pocos muestreos se han cogido datos suficientes del ecosistema. En la mayoría de los casos apenas los únicos datos que se cogen, son el tipo de ecosistema, por ejemplo: aguas de la costa, glaciar, tierra, … lo cual nos va a dar varios problemas en cualquiera de los casos.

Los datos de cada ecosistema son muy importantes ya que numerosos microorganismos pueden vivir en un medio muy concreto y con características físico-químicas muy concretas. El hecho que numerosos ecosistemas tengan una organización muy cerrada, podría indicar que la ayuda mutua entre diversas especies da la posibilidad de que una especie pueda vivir en un entrono en el que sola le seria imposible. Por ello cuanto mas datos se hayan recaudado la clasificación será mas

completa.

Por otra parte está la cantidad de secuencias que se describen en estas muestras. Ya que en estos últimos años se ha visto como la cantidad de datos biológicos se han ido multiplicando, sobretodo desde la aparición de las nuevas tecnologías de secuenciación masiva. Esto ha llevado a desarrollar las técnicas de metagenómica con las en el día de hoy se analizan ecosistemas enteros, y se calculan las redes de interacción entre las especies que se encuentran en ellos. Para trabajar con estos datos hay que desarrollar nuevas formas de almacenar la información de forma que resulte fácil analizarla después, y poder sacar nuevas conclusiones.

A la hora de analizar secuencias con el fin de clasificar filogenéticamente las secuencias de bacterias que se hayan encontrado en los diversos ecosistemas, se recurre a las secuencias referentes al gen del 16S RNA [6]. Con ello se han creado diversas bases de datos en las que las bacterias se clasifican taxonómicamente en base a este gen. Estas bases de datos también suelen tener datos referentes al lugar donde se hayan recogido las muestras [7]. Para estos casos en los que es necesario desarrollar tecnologías nuevas para el análisis de las secuencias desde bases de datos relacionales, desde las cuales se pueden analizar filogenéticamente [8]. Estos programas suelen trabajar de forma muy concreta para una única solución, como describir la taxonomía de ciertas comunidades ya descritas en otras bases de datos [9].

Por otra parte también se han creado aplicaciones para trabajar con bases de datos ya existentes y analizar la información que haya en ellos [10]. Estos programas realizan una búsqueda de los datos que se les pida en ciertas bases de datos y después guarda esa información en una base de datos propia, sobre la cual posteriormente se realizaran los análisis pertinentes.

En este caso lo que se pretende hacer mediante este trabajo son los siguientes puntos:

- Crear nuestro propio sistema de descargas de ficheros desde la base de datos del Genbank, donde se encuentran los archivos con información referente a las secuencias del 16S.

- Crear nuestra propia base de datos relacional, en la que guardaran de forma ordenada todos los datos referentes a cada entrada que se vayan logrando a partir de los programas que se vayan corriendo para lograr mas información referente a dichas entradas.

- Organizar los programas de tal forma que después de que la base de datos este cargada, se pueda actualizar sin borrar los datos ya existentes en ella. De la misma forma que todos corran de forma ordenada para que no se colapse ni de problemas a la hora de cargar la base de datos.

– Los programas de análisis que se van a correr, después de tener la base de datos cargada, nos van a permitir analizar las siguientes características de las entradas:

1. La evolución de las secuencias y de las especies a lo largo de los años en la base de datos del Genbank

2. La cantidad de taxones descritos en la base de datos, y todos los tipos existentes de los mismos

3. La relación de los medios con las especies que se encuentran en ellos y la relación de las diferentes especies entre ellas mismas dentro de cada ecosistema descrito.

Para el buen uso de dicho programa es necesario tener primero todos los programas y módulos ya existentes instalados (ver anexos). Ya que la falta de uno de estos requisitos provocaría el mal funcionamiento de todo el proceso. Esto es importante, ya que aparte de los scrips que se han creado para que este paquete de programas funcione, se utilizan varios programas extra, que son necesarios para que todas los programas funcionen correctamente.

Por otra parte este paquete tiene que estar instalado en una computadora con bastante capacidad, ya que las nuevas tecnologías de secuenciación masiva están creando metagenomas enteros, los cuales tienen numerosas nuevas secuencias y estas se traducen en mas espacio cada vez que se actualiza la base de datos. Estos avances están provocando que bases de datos como la del Genbank, cada vez esté mucho mas llena de metagenomas enteros. Cada vez que se actualiza la base de datos del Genbank la cantidad de secuencias aumenta cada vez mas. Por ello las entradas cada vez son mas completas, ya que se realizan metagenomas de diferentes muestras y cada vez son mas completos. Por está razón las entradas han ido evolucionando paulatinamente, ya que las primeras entradas apenas eran de bacterias cultivables en laboratorio y las de hoy en día han podido ser extraídas prácticamente en cualquier lugar del planeta. Pero esto a la hora de crear un parser nos provoca los siguientes problemas: de las bacterias cultivables del principio el único dato respecto al origen de la bacteria es que es una bacteria cultivada. Por otro lado existen numerosas entradas en las que ni siquiera existen datos si la bacteria ha sido cultivada o no, o si ha sido extraída de algún medio. Por otro lado en los casos en los que aparecen datos referentes al lugar de extracción de la muestra no tienen un criterio estandarizado, lo cual produce muchos errores a la hora de conseguir los datos. El hecho que no exista un modelo estandarizado nos lleva a tener referencias muy vagas del entorno, por ejemplo: agua marina, tierra, agua, … referencias que no nos dicen gran cosa. Sin embargo en los últimos años, si que los datos referentes al medio de extracción son muchos mas completos, comentan desde el tipo de hospedador, país o coordenadas donde fue obtenida la muestra. Con estos datos se puede realizar un análisis mucho mas exhaustivo ya que aparte de ponerte el medio de obtención, también te ponen el país y las coordenadas, lo cual facilita el análisis de diferentes medios ecológicos.

Todos estas variables hay que tenerlas en cuenta a la hora de realizar el parser, puesto que olvidar uno de estos datos podría acarrear problemas de identidad de la muestra y como resultado se clasificaría mal.

La necesidad de crear una base de datos relacional tan completa es precisamente por la simple razón de tener que analizar múltiples campos de forma muy diversa:

- El campo referente al medio en el que se haya extraído la muestra es necesario sacar toda la información posible de ello. Ya que la falta de información en este campo es bastante habitual, y para el posterior análisis de los ecosistemas donde se hayan extraído las muestras, es de gran importancia poder tener el medio lo mas descrito posible.
-
- A la hora de analizar las secuencias de la base de datos para asignar un taxón a cada una de ellas, hay que tener en cuenta varios valores, para la hora de asignar una especie a cada secuencia. El problema es que en la base de datos existen múltiples entradas de secuencias que hacen referencia a la misma especie. Por ello primero hay que clusterizar las secuencias en base al parámetro de diferenciación de especies para las bacterias (una similitud superior al 97% se considerará la misma especie). Aunque este valor puede ser alterado en caso de que la nueva bibliografía que se vaya publicando indique otro valor.

Todos estos valores hay que tratarlos con mucha atención ya que una clasificación errónea puede cambiar por completo el análisis final. Por ello la metodología es importante a la hora de llevarla a cabo.

Métodos:

+Soporte informático:

Para realizar el siguiente trabajo se han utilizado el siguiente soporte informático:

- El sistema operativo utilizado a sido Linux:3.0.0-25-generic Ubuntu11.10. Sobre este soporte se han corrido todos los scrips que se han creado.

- El lenguaje de programación a sido Perl: v5.12.4. Con él se han realizado todos los scrips que se han utilizado.

- Para realizar los cálculos estadísticos y los resultado en gráficos se ha utilizado R: 2.15.1.

- A la hora de crear la base de datos se ha utilizado PostgreSQL: 9.1.5.

Una vez que se disponía de dicho soporte, se pasó a conseguir los datos con los que se iba a trabajar y a crear los scrips y la base de datos que serian necesarios, para el posterior análisis de la información lograda.

+ La base de datos:

Para crear la base de datos se ha llevado a cabo mediante el programa postgresql. Con este programa se ha creado la estructura de la base de datos, la cual está descrita en el archivo bdmicro.sql (ver anexos). De esta forma la base de datos ha quedado con esta estructura:

Mediante esta estructura los datos se distribuyen de forma que el análisis de cada una de las características sea mas fácil de llevar a cabo:

La tabla micro: En esta tabla se vuelcan todo los datos del Genbank logrados en el primer scrip. Son datos brutos y sin tratar, simplemente lo que devuelve el parser.

La tabla de muestra: Las entradas que tengan el mismo titulo y los mismos autores se consideraran de la misma muestra. Esto se lleva a cabo después de la corrección de estos, como ya veremos posteriormente.

La tabla de ambientes: En dicha tabla se guardarán todos los datos referentes al medio de extracción en el que se hayan logrado las muestras. En este caso, para cada ambiente diferente, se le asignará un numero arbitrario.

La tabla de secuencias: En esta tabla las secuencias son agrupadas en clusteres, que se tomarán como referentes para analizar cada caso, con los cuales se trabajará mas adelante.

La tabla de especies: mediante la clusterización de las secuencias se le asigna un taxón a la secuencia representante de cada conjunto.

La tabla de taxones: se guardan todos los datos taxonómicos de la secuencia representante de cada cluster logrados mediante un blast. El taxón asignado a cada caso y la puntuación con la que se le ha dado.

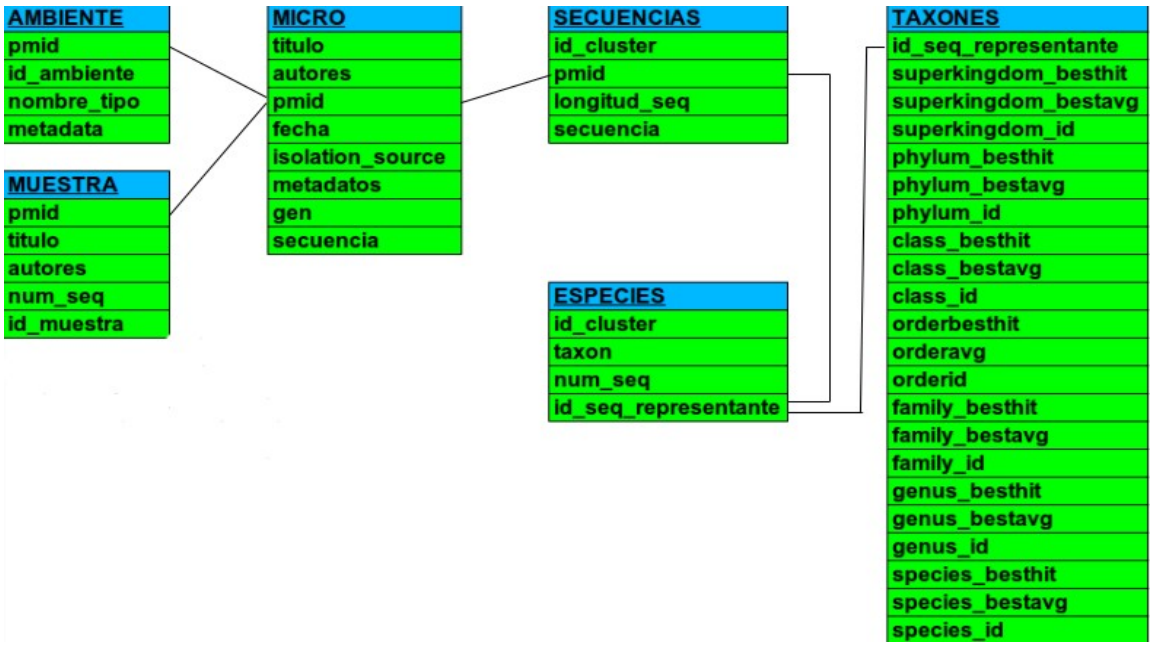


Imagen.1: Esquema de la estructura de la base de datos. Los campos que aparecen unidos con las lineas son los campos que están interrelacionados.

La estructura de esta base de datos está repartida en dos como se puede ver (imagen.1); por un lado las tablas que hacen referencia a las muestras y el medio en el que fue extraída cada muestra. Por otro lado están todos los datos referentes a la secuencias y a la taxonomía referente a cada entrada. De esta forma la búsqueda de la información está mucho mas agrupada y mas fácil de manejar para cuando se tenga que lograr datos desde las tablas.

A este esquema se le pueden sumar otras tablas de querer añadir mas datos referentes a las entradas. Como la revista de publicación, o en los casos que aparecen , los datos referentes al tipo de clon al que hace referencia la entrada.

+ Programas para rellenar la base de datos

Una vez creada la base de datos se llevo a cabo la introducción de los datos en la misma. Los datos que se han utilizado se han logrado de la base de datos del Genbank. Concretamente los archivos se han descargado de esta dirección: ftp://ftp.ncbi.nih.gov/genbank/. De dicha lista de ficheros solo se descargaron los de la división gbenv, donde nos encontraremos las secuencias relacionadas con el 16S y otras secuencias marcadoras relacionadas con dicha secuencia cromosómica.

La descarga se llevo a cabo con el programa wget con la siguiente linea de comandos: wget ftp://ftp.ncbi.nih.gov/genbank/gbenv*. Una vez logrado todos estos archivos, se guardaron en el directorio que iba a funcionar como lugar de trabajo.

Una vez realizada la descarga, todos estos archivos son descomprimidos y

analizados por el programa de microparser.pl (ver anexos II). Este programa lee uno a uno todo los archivos e introduce en la base de datos los datos necesarios para llevar a cabo el posterior análisis.

Este programa extrae de los ficheros los siguientes campos de cada entrada: titulo, autores, numero de identificador del Genbank, la fecha de la entrada, el gen que codifica (solo nos quedamos con las secuencias del 16S), la secuencia del gen, el medio en el que se extrajo la muestra (si aparece), datos referentes al medio en el que se extrajo la muestra (si aparece). Esto que en un principio parece una tarea facil, se ve complicada con diversos tipos de entradas, así como la falta de ciertos campos en muchas entradas, lo cual deja información sin completar para ciertas secuencias.

```
LOCUS       EU748459             1360 bp    DNA     linear   ENV 29-SEP-2008
DEFINITION  Uncultured bacterium clone hoa61_09c09 16S ribosomal RNA gene,
            partial sequence.
ACCESSION   EU748459
VERSION     EU748459.1  GI:190403597
KEYWORDS    ENV.
SOURCE      uncultured bacterium
  ORGANISM  uncultured bacterium
            Bacteria; environmental samples.
REFERENCE   1  (bases 1 to 1360)
  AUTHORS   Godoy-Vitorino,F., Ley,R.E., Gao,Z., Pei,Z., Ortiz-Zuazaga,H.,
            Pericchi,L.R., Garcia-Amado,M.A., Michelangeli,F., Blaser,M.J.,
            Gordon,J.I. and Dominguez-Bello,M.G.
  TITLE     Bacterial community in the crop of the hoatzin, a neotropical
            folivorous flying bird
  JOURNAL   Appl. Environ. Microbiol. 74 (19), 5905-5912 (2008)
   PUBMED   18689523
REFERENCE   2  (bases 1 to 1360)
  AUTHORS   Godoy-Vitorino,F., Ley,R.E., Gao,Z., Pei,Z., Ortiz-Zuazaga,H.,
            Pericchi,L.R., Garcia-Amado,M.A., Michelangeli,F., Blaser,M.J.,
            Gordon,J.I. and Dominguez-Bello,M.G.
  TITLE     Direct Submission
  JOURNAL   Submitted (27-MAY-2008) Biology, University of Puerto Rico, Rio
            Piedras Campus, PO Box 23360, San Juan, Puerto Rico 00931, USA
FEATURES             Location/Qualifiers
     source          1..1360
                     /organism="uncultured bacterium"
                     /mol_type="genomic DNA"
                     /isolation_source="adult hoatzin crop"
                     /host="Opisthocomus hoazin"
                     /db_xref="taxon:77133"
                     /clone="hoa61_09c09"
                     /environmental_sample
                     /country="Venezuela"
                     /PCR_primers="fwd_name: 8F, fwd_seq: agagtttgatymtggctcag,
                     rev_name: 1513R, rev_seq: tacggytaccttgttacgactt"
     rRNA            <1..>1360
                     /product="16S ribosomal RNA"
ORIGIN
        1 gatgaacgct agctacaggc ttaacacatg caagtcgagg ggaaacgacg gcggggtttc
       61 ggccttgccg ggcgtcgacc ggcggatggg tgagtaacgc gtatccaacc tgcccctgtc
```

Imagen.2: Tipo de entrada mas común del Genbank.

En la imagen de la entrada mas común del Genbank podemos ver los tipos de campos de los que hay que extraer la información. En esta entrada, se ven practicamente todos los campos posibles de extraer, ya que en otras muchas entradas faltan varios campos (isolation_source, host, country, Title,...). Otro tipo de entrada bastante común en el Genbank es la siguiente:

```
LOCUS       EU808050               789 bp    DNA     linear   ENV 29-JUN-2008
DEFINITION  Uncultured bacterium clone Chlplus_CL-030610_OTU-1 16S ribosomal
            RNA gene, partial sequence.
ACCESSION   EU808050
VERSION     EU808050.1  GI:192786864
KEYWORDS    ENV.
SOURCE      uncultured bacterium
  ORGANISM  uncultured bacterium
            Bacteria; environmental samples.
REFERENCE   1  (bases 1 to 789)
  AUTHORS   Noguera,D.R., Yilmaz,L.S., Harrington,G. and Goel,R.C.
  JOURNAL   (in) IDENTIFICATION OF HETEROTROPHIC BACTERIA THAT COLONIZE
            CHLORAMINATED DRINKING WATER DISTRIBUTION SYSTEMS. AWWA Research
            Foundation, 6666 West Quincy Avenue, Denver, CO, USA (2008), In
            press
REFERENCE   2  (bases 1 to 789)
  AUTHORS   Noguera,D.R., Yilmaz,L.S., Harrington,G. and Goel,R.C.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-JUN-2008) Department of Civil and Environmental
            Engineering, University of Wisconsin - Madison, 1415 Engineering
            Dr., 3207 Engineering Hall, Madison, WI 53705, USA
FEATURES             Location/Qualifiers
     source          1..789
                     /organism="uncultured bacterium"
                     /mol_type="genomic DNA"
                     /isolation_source="chloraminated bench-scale chemostat"
                     /db_xref="taxon:77133"
                     /clone="Chlplus_CL-030610_OTU-1"
                     /environmental_sample
     rRNA            <1..>789
                     /product="16S ribosomal RNA"
ORIGIN
        1 tcgtggggca gcgcaggtag caatactggg cggcgaccgg caaacgggtg cggaacacgt
       61 acacaacctt ccgagaagtg gggaatagcc cagagaaatt tggattaata ccccgtaaca
      121 taacgatgtg gcatcacatt gttattatag cttcggcgct tcttgatggg tgtgcggctg
      181 attagatagt tggcggggta acggcccacc aagtctacga tcagtagctg atgtgagagc
      241 atgatcagcc acacgggcac tgagacacgg gcccgactcc tacgggaggc agcagtaagg
```

Imagen.3: Un tipo de entrada bastante habitual en el Genbank, en el que el apartado del titulo está dentro del apartado Journal seguido del limitador (in).

En este otro tipo de entrada, no encontramos con el siguiente problema, el titulo no está limitado de ninguna manera universal. En múltiples entradas el titulo está escrito en mayúsculas, pero en otras no, en otras entradas está limitado por un punto al final (.), pero en otras no. Además, este ultimo limitador no es muy eficaz a la hora de ponerlo como fina del titulo, ya que varios títulos tienen números referentes a distancias, concentraciones, nombres de abreviaciones, etc... que en sí llevan el símbolo del punto. En otros casos su limitante es el punto y coma (;). Y de esta manera tan dispar quedan limitados ciertos títulos.

Por otro lado también existen entradas en las que ni aparece titulo ni datos sobre donde se a extraído la muestra. En estos últimos casos la información queda bastante pobre, ya que no se va a poder delimitar como es debido la muestra.

Por otro lado también existen otros tipos de entradas las cuales no siguen ningún formato ya descrito hasta ahora. Pero a ser muy pocas entradas, no nos hemos preocupado de realizar una parser especifico para ellas. Estas entradas que no se han conseguido parsear adecuadamente apenas son 394. Al tener 3.349.676 entradas, ese numero no nos parecía relevante para realizar primero la búsqueda dentro de los ficheros del Genbank, y la posterior preparación de los parser requeriría mas esfuerzo que la información que estas nos puedan aportar.

Aunque se puede conseguir mas información de cada entrada, como si pertenece a un clon en concreto o, en los casos que aparece, la numeración del taxón, o donde fue publicado, etc … no nos parecen importantes a la hora de extraer los datos, por las siguientes razones:

- La revista donde fue publicada nos es indiferente, ya que ni hace referencia a ningún tipo de nicho ecológico ni ayuda a clasificar taxonómicamente.
- El numero de taxón nos es indiferente, ya que en varias entradas no aparece, y en muchas en las que aparece no nos podemos fiar que esté bien identificado.
- Los clones de ciertas especies de microorganismos, se catalogarán posteriormente como la misma especie, mediante el programa CD-HIT-EST, dentro del mismo cluster que hará referencia a una especie.

Por otro lado en estos archivos hay múltiples entradas que hacen referencia a secuencias que interaccionan con el 16S de ciertas especies, pero sin llegar a ser la propia secuencia del 16S. Por ello para quitar las secuencias que no sean del propio 16S se llevan a cabo dos filtros por entrada en el programa microparser.pl:

-En el caso del gen: solo nos quedamos con los que hacen referencia a la secuencia 16S del ribosoma.

-En el caso de la secuencia: solo nos quedamos con las secuencias cuya longitud comprenda entre 200 y 1.800 nucleótidos. En base a que son los limites establecidos para los 16S conocidos.

Por otra parte este programa crea otros 3 archivos, los cuales servirán para los trabajos que se van a realizar para lograr mas información de cada entrada.:

-Un archivo fasta con todas las secuencias del 16S, que se utilizará para hacer el blast posteriormente, y de esta forma asignar una especie a cada secuencia que se haya extraído.

-Un archivo con todos los títulos y los autores separados por pipe (|). Gracias a este archivo se llevarán a cabo las correcciones de los autores y los títulos existentes en los datos extraídos del Genbank y se agruparán posteriormente en diversas muestras.

-Un Archivo con el identificador del Genbank, el titulo y los autores de cada entrada separadas por un pipe (|) cada uno. Con el fin de asignar posteriormente el numero de muestra a cada caso, mas tarde un programa asignará una muestra a un titulo y autores después de haberlos corregidos, mediante el archivo anterior.

El segundo archivo se crea para llevar a cabo una corrección de los mismos. Ya que se han encontrado numerosos fallos para algunos títulos y autores, como diferencias entre ellos perteneciendo al mismo articulo, es decir, tendrían que ser iguales. Estos fallos van desde errores ortográficos, cambios de mayúsculas a minúsculas y viceversa, desapariciones de espacios, cambio de orden de los nombres, etc. Mediante otros programas se quiere lograr solucionar los fallos y poder determinar de esta manera cada muestra de una forma mas fiable.

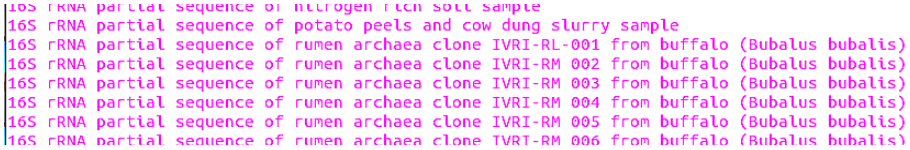


Imagen.4: En este caso vemos como un titulo similar puede tener diferentes entradas.

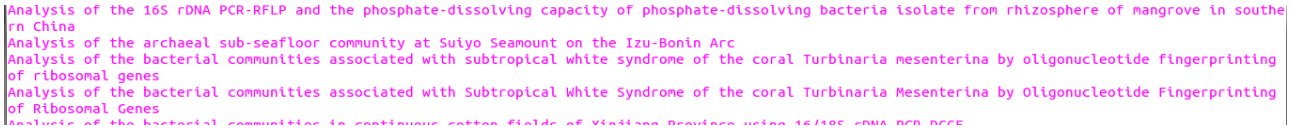


Imagen.5: En este caso vemos como siendo el mismo titulo tenemos diferentes entradas, por tener letras cambiadas de orden de posición o de tipo (mayúsculas/minúsculas), además de los errores ortográficos que tiene algunas entradas.

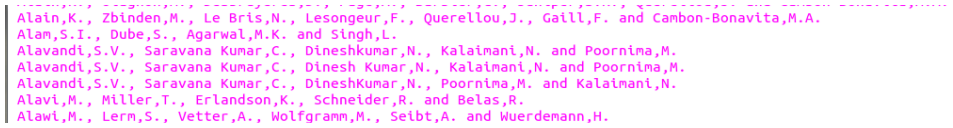


Imagen.6: En esta imagen se puede ver como siendo los mismos autores en cada entrada tienen diferencias, pese a ser la misma lista. Estas diferencias pueden ir desde cambio de orden de los nombres, hasta tener algún nombre mal escrito.

Para realizar estas correcciones aplicamos este modulo de Perl: LevenshteinXS. Con este modulo podemos definir cuantas diferencias puede existir entre dos objetos y asumirlo como igual o diferente. En nuestro caso ponemos el filtro en 13 diferencias para los títulos y autores a la vez. De esta forma cualquier titulo y autor que tenga menos diferencias que las definidas con otro titulo, se consideraran iguales, y se sustituirá el titulo analizado por el primero que se haya definido. En otros palabras, se clusteriza los títulos y los autores y se escoge uno como representante.

El trabajo de solucionar las diferencias entre los títulos y los autores lo llevan a cabo los siguientes programas:

El programa idmuestra.pl (ver anexos II) coge el archivo de autores y títulos que ha creado el scrip microparser.pl, y mediante el modulo LevenshteinXS elimina la mayoría de los errores existentes. Para ello, primero ordena todas las entradas por orden alfabético (con la orden sort de UNIX) y posteriormente eliminando todas las

repeticiones (mediante la orden uniq de UNIX). Mediante estas simples ordenes eliminamos una gran cantidad de trabajo a la computadora a la hora de determinar las diferencias entre las cadenas que se están comparando mediante el modulo LevenshteinXS. Al final este programa devuelve un archivo de títulos y autores, ya con todos los errores corregidos, con los cuales realizaremos las asignación del tipo de muestra en el siguiente paso. Para finalizar, con el fin de evitar repetir varias veces la misma búsqueda, después de la corrección de los errores, se realiza otro sort y otro uniq sobre el archivo resultante, de esta forma el análisis que se realizará en el siguiente paso será mucho mas eficaz.

En este caso los autores y títulos corregidos se guardan en la tabla de 'muestra'. Aun así los titulo originales se quedan guardados en la tabla 'micro', ya que se puede sacar información de ellos. Después de las correcciones se pueden perder información importante; como puede ser el caso de los trabajos realizados con diferentes cepas (imagen.4). Por ello, en un futuro si se quiere llevar a cabo un proceso de text-mining se podría sacar información a partir de estas entradas. Esto vendrá bien en caso de querer sacar mas información respecto al medio en el que se extrajo la muestra. Ya que la información de este campo es bastante pobre.

La asignación de muestra se basa en que si un conjunto de entradas de la base de datos tiene los mismos títulos y los mismos autores, se consideran la misma muestra. Esta asignación es arbitraria, simplemente poniendo el un numero de muestra a cada conjunto de entradas. En este caso, también se cuenta cuantas secuencias existen en cada muestra, de esta forma también nos podremos hacer una idea de la diversidad analizada en cada trabajo.

El interés de agrupar las entradas de esta forma reside en facilitar el análisis posterior de cada medio. Cada una de estas muestras está relacionada con un trabajo, artículo publicado, en concreto. En la mayoría de los casos estos trabajos se realizan sobre un solo medio. Aunque en ocasiones existen ciertos trabajos que recoge muestras de diferentes medios (sobre todo los relacionados con el medio acuático). Por ello la primera clusterización se lleva a cabo en base a los artículos y posteriormente, en los casos que existen, también se tiene en cuenta el medio en el que se ha realizado el aislamiento de la muestra. Esto facilitará los análisis posteriores de los ambientes en los que se haya conseguido cada muestra.

A la hora de agrupar los datos en la tabla de 'muestra' de la base de datos, lo llevamos a cabo mediante el programa muestratabla.pl (ver anexos II). Mediante este programa se buscan todos titulo y autores que hagan referencia a los que ya estén corregidos, y extrae el identificador de cada uno y les asigna un numero de muestra, así como la cantidad de secuencias que existen en cada muestra. Gracias a estos datos a la hora de llamar por ambientes y muestras la búsqueda queda mucho mas reducida y será mucho mas eficaz.

Por otro lado también se completa la tabla de 'ambientes' mediante el programa metadata_ambiente.pl (ver anexos II). Donde para cada muestra se añadirán los diferentes datos referentes a la extracción de la misma. En varios casos a la que hemos considerado la misma muestra podrá tener diferentes tipos de

identificadores de ambientes. Podría ser el caso de un estudio realizado a lo largo de un río donde la muestra se referiría a todo el estudio realizado sobre el mismo río, y el identificador de ambiente haría referencia a las distintas tomas realizadas a los largo del transecto del río.

En esta ultima tabla que haría referencia a los tipos de ambientes en los que se ha extraído la muestra nos va a servir el ultima instancia a dividir las comunidades bacterianas por medios ambientales. Aunque el hecho de que en la mayoría de los casos falte mucha información respecto a estos campos va a dejar fuera de la posibilidad de analizar a muchas secuencias y especies.

+ La asignación de las especies:

Por otro lado, mediante el fasta creado a partir de primer scrip (microparser.pl) se van a lograr los taxones de cada secuencia. Para ello, se va a seguir un proceso de eliminación de redundancia de secuencias, clusterización de las mismas y la posterior clasificación taxonómica de todas ellas.

El primer paso a llevar a cabo, es la eliminación de la redundancia en la secuencias y la custerización de estas. Para ello utilizaremos el programa CD-HIT-EST con el que las secuencias serán clusterizadas, y por cada no de ellos se selecciona una secuencia representante con la que se trabajará posteriormente, en vez de trabajar con todas las secuencias.

El comando del programa CD-HIT-EST utilizado es el siguiente:

```
$ cd-hit-est -i fasta.fa -o outfasta -l 200 -c 0.97 -M 4000 -aL 0.8 -aS 0.8 -r 1
```

Mediante este comando solo se trabajará con las secuencias mayores de 200 nucleótidos y se asignará 4000 Megas de memoria para que el programa pueda correr sin ningún problema. Por otro lado, las secuencias se agruparán en base a los siguientes criterios: las secuencias tienen que ser parecidas en un 97% y tienen que tener un coverage del 80% tanto en la cadena larga como en la corta a la hora de realizar la comparación. Esta comparación se llevará a cabo en las dos direcciones de la cadena, ya que no se sabe a ciencia cierta en que dirección se ha secuenciado el rRNA.

Una vez corrido el programa CD-HIT-EST, nos devuelve tres archivos con el nombre base outfasta:

- outfasta: Es el archivo de secuencias fasta solo con las secuencias representante de cada cluster. De esta forma a la hora de realizar al blast solo se realizará en sobre este archivo, y así será mucho mas rápido ya que cada secuencia de este fasta en la base de datos se relaciona con todas las que so similares a la suya.

– Outfasta.clstr: Es el archivo el que aparecen los clusteres con todos los numeros de identificadores que existe en cada uno de ellos, y entre ellos estaría marcado con un asterisco (*) la secuencia representante de cada cluster. Con este archivo se relacionan todas las secuencias de cada cluster con la secuencia representante, con la que posteriormente se realizará el blast.

– outfasta.bak.clstr: Este archivo es igual que el .clstr con la diferencia que este archivo comienza con una columna en la que especifica el numero de cluster de cada secuencia.

Una vez logrado estos archivos, sus datos van a ser introducidos en la base de datos en la tabla "secuencias", mediante el programa cdhitparser.pl (ver anexos). Este programa simplemente coge el archivo .clstr y lo parsea sacando de el los siguientes datos e introduciéndolos en la tabla de "secuencias": el identificador del Genbank, el identificador del cluster, la longitud de la secuencia y la secuencia. Esta ultima la saca de la tabla de "micro". Por otro lado también carga la tabla de especies introduciendo el identificador de cluster, la cantidad de secuencias que hay en cada uno de ellos y la secuencias representante.

Posteriormente con el archivo fasta que se ha logrado mediante el CD-HIT-EST (el archivo outfasta), se va a realizar un blast con el que se asignará el taxón correspondiente a cada secuencia. Para realizar esta asignación es necesario tener instalado en el ordenador el blast (blastall). De la misma forma, también es necesario tener las bases de datos de greengenes con la que se van a realizar las asignaciones taxonómicas:

El archivo que vamos a utilizar como base de datos lo descargamos desde la siguiente dirección, donde aparecen todas las secuencias no alineadas de las procariotas que se hayan descrito en la base de datos del greengenes: greengenes.lbl.gov/Download/Sequence_Data/Fasta_data_files/current_prokMSA_unaligned.fasta.gz

Después de realizar la descarga el archivo se descomprime y se le da el formato deseado con el programa mtax.pl (ver anexos I). Este programa ya estaba creado y funcionaba junto con el programa taxbuild_NOT_EUK.pm (ver anexos I), que también estaba ya creado. De esta forma, mediante este comando se le da formato a la base de datos del greengenes de forma que el programa de blastall pueda interpretarlo.

```
$ perl mtax.pl
```

El programa mtax.pl (ver anexos I) tiene que ser modificado cada vez que se introduzca en un nuevo ordenador, ya que sus llamadas son mediante rutas absolutas a programas concretos.

Por otra parte este programa, necesita de otro programa ya existente para trabajar con él. Este programa, taxbuild_NOT_EUK.pm (ver anexos I) trabaja con los archivo de la base de datos Taxonomy del NCBI, que hace referencia a la base de datos de las bacterias, como podemos ver en la pagina en esta dirección:

ftp://ftp.ncbi.nih.gov/pub/taxonomy/

Mediante estos archivos, se realiza una relación desde el nivel de especie hasta el nivel de reino. En otras palabras, en base al nombre de la especie, se busca el nodo que identifica a su genero, y del nodo del genero se busca al nodo de la familia y así sucesivamente hasta llegar a nivel de reino, como se puede ver en la imagen (Imagen.7):

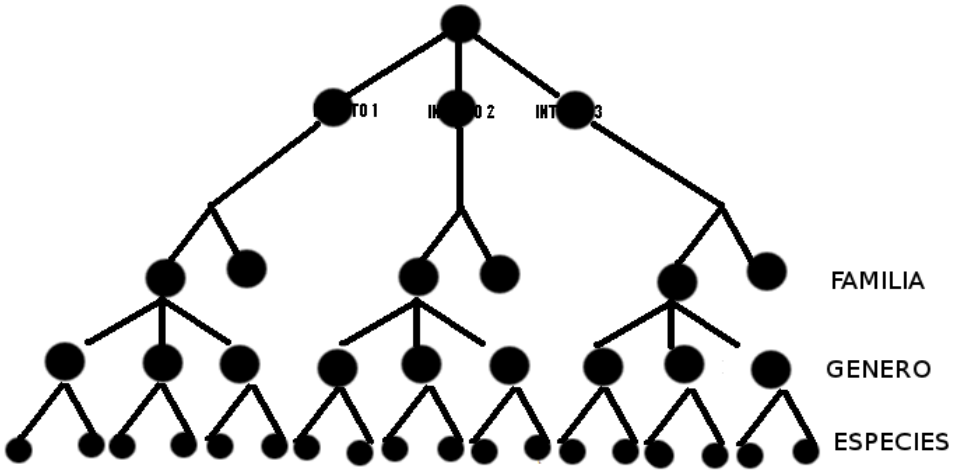


Imagen.7. Estructura de nodos del greengenes. En base al nombre de la especie va estructurando el árbol taxonómico hasta llegar a nivel de reino, en este caso, el de las bacterias.

De esta forma para cada especie se guarda una estructura taxonómica concreta y una numeración para hacer mas fácil la comparación y la llamada a esta. Al final lo que devuelve es una archivo con el nombre greengenes.tax, en el que se encuentran todas las filogenias conocidas de todas las especies, que han sido descritas en dicha base de datos. Este archivo tiene todas las especies organizadas por arboles taxonómicos a las cuales a cada una se le asigna un numero de identificador de especie:

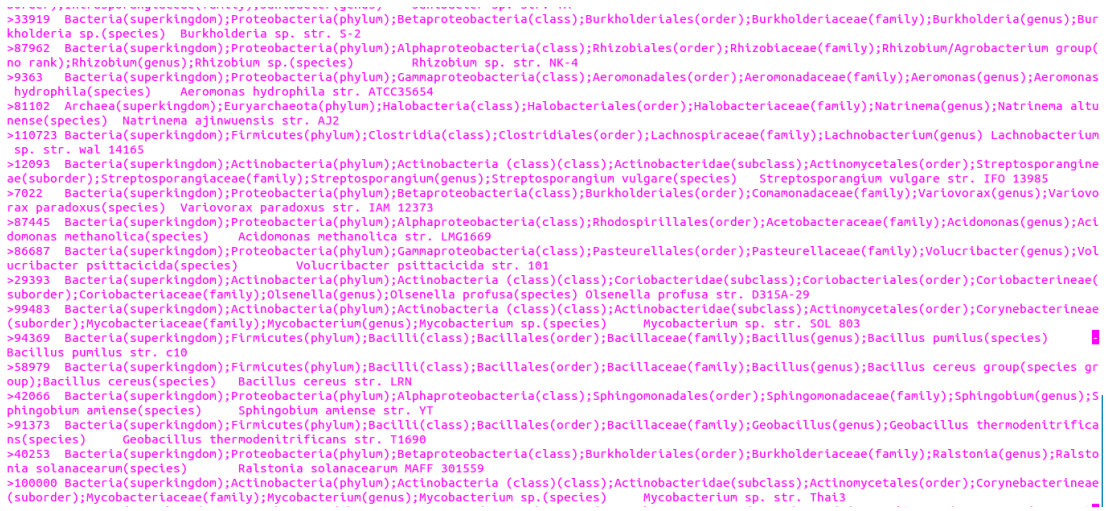


Imagen.8: formato del archivo de greengenes.tax, en el que se ve la estructura de asignación de valores taxonómicos a cada especie, así como la asignación de un numero identificativo para cada especie.

Posteriormente, para que el blast pueda usar este archivo como base de datos para lanzar contra él las secuencias que estamos analizando, hay que dar el formato apropiado para que lo admita el programa blastall. Para ello se utiliza el siguiente comando:

$ formatdb -i greengenes.tax -p F

Una vez que todo los archivos tienen el formato que admite el blastall, se lleva a cabo el siguiente comando para fraccionar el archivo fasta que nos ha devuelto el CD-HIT-EST para que cada archivo que se le pase al blast sea de un tamaño adecuado y no se bloquee el programa. En este caso se crean archivos de 50.000 secuencias cada uno:

$ perl conteofasta.pl

De esta forma el conteofasta.pl (ver anexos II) divide el archivo del fasta que nos creo el CD-HIT-EST, ya que puede llegar a ser un archivo de demasiadas secuencias para realizar el blast. Así, lo que iba a ser un gran problema para el blastall queda solucionado, facilitando el poder computacional del ordenador en el que se este corriendo el programa.

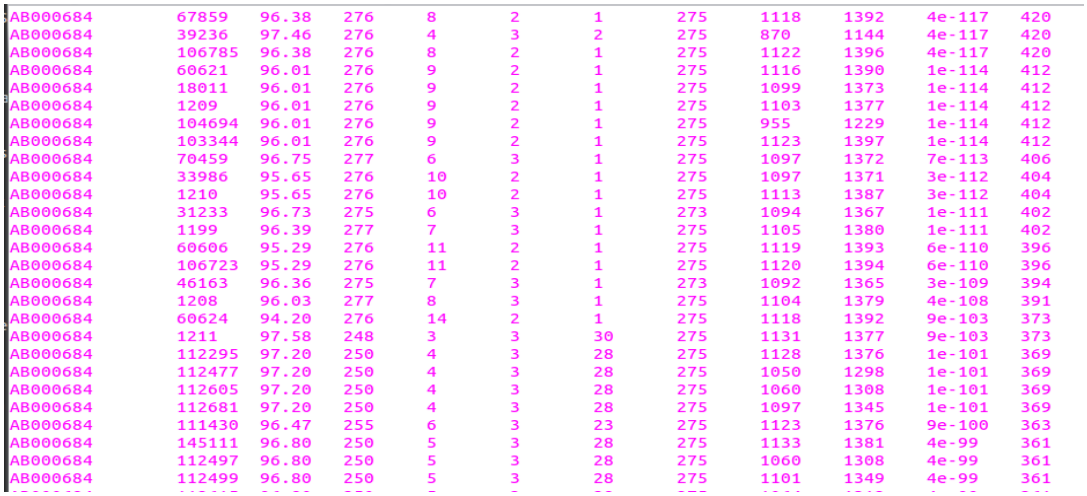
Una vez que el fasta ya esté fraccionado, se realizará el blast sobre los archivos que haya devuelto el programa anterior mediante el siguiente comando:

$ blastall -p blastn -i fast*.txt -d greengenes -o file*.fas.blast -m 8 -e 1e-03 -D 200 -a 2

Con este comando se consigue lo siguiente: mediante la base de datos greengenes se logrará un archivo con el formato .fas.blastn y con el nombre file en el que los hits que se logren tendrán como mínimo un e-value 1e-03 con un máximo de 200 hits por secuencia, utilizando dos núcleos de la computadora. Se utiliza el proceso blastn ya que se utiliza nucleótidos contra nucleótidos a la hora de hacer la comparación.

Este comando esta dentro del programa blastn.pl (ver anexos III), ya que es necesario un bucle para leer todos los archivos y que devuelva los resultados de una forma coherente y sin problemas.

El formato .fas.blastn que devuelve el blastall tiene esta forma:



AB000684	67859	96.38	276	8	2	1	275	1118	1392	4e-117	420
AB000684	39236	97.46	276	4	3	2	275	870	1144	4e-117	420
AB000684	106785	96.38	276	8	2	1	275	1122	1396	4e-117	420
AB000684	60621	96.01	276	9	2	1	275	1116	1390	1e-114	412
AB000684	18011	96.01	276	9	2	1	275	1099	1373	1e-114	412
AB000684	1209	96.01	276	9	2	1	275	1103	1377	1e-114	412
AB000684	104694	96.01	276	9	2	1	275	955	1229	1e-114	412
AB000684	103344	96.01	276	9	2	1	275	1123	1397	1e-114	412
AB000684	70459	96.75	277	6	3	1	275	1097	1372	7e-113	406
AB000684	33986	95.65	276	10	2	1	275	1097	1371	3e-112	404
AB000684	1210	95.65	276	10	2	1	275	1113	1387	3e-112	404
AB000684	31233	96.73	275	6	3	1	273	1094	1367	1e-111	402
AB000684	1199	96.39	277	7	3	1	275	1105	1380	1e-111	402
AB000684	60606	95.29	276	11	2	1	275	1119	1393	6e-110	396
AB000684	106723	95.29	276	11	2	1	275	1120	1394	6e-110	396
AB000684	46163	96.36	275	7	3	1	273	1092	1365	3e-109	394
AB000684	1208	96.03	277	8	3	1	275	1104	1379	4e-108	391
AB000684	60624	94.20	276	14	2	1	275	1118	1392	9e-103	373
AB000684	1211	97.58	248	3	3	30	275	1131	1377	9e-103	373
AB000684	112295	97.20	250	4	3	28	275	1128	1376	1e-101	369
AB000684	112477	97.20	250	4	3	28	275	1050	1298	1e-101	369
AB000684	112605	97.20	250	4	3	28	275	1060	1308	1e-101	369
AB000684	112681	97.20	250	4	3	28	275	1097	1345	1e-101	369
AB000684	111430	96.47	255	6	3	23	275	1123	1376	9e-100	363
AB000684	145111	96.80	250	5	3	28	275	1133	1381	4e-99	361
AB000684	112497	96.80	250	5	3	28	275	1060	1308	4e-99	361
AB000684	112499	96.80	250	5	3	28	275	1101	1349	4e-99	361

Imagen.9: En esta imagen se ve un ejemplo del resultado del blast. Donde los valores que nos interesan son la primera columna (secuencia que se está analizando), la segunda (la numeración de la especie de greengenes), la tercera (el porcentaje de identidad) y la undécima (e-value).

Una vez que haya acabado de correr el blast, uniremos todos los archivos de resultados en uno solo con el que posteriormente correremos el programa asigna16S.pl. Para ello realizamos el siguiente comando:

$ cat *.fas.blastn > fichero.fas.blastn

Utilizando el archivo que este haya devuelto, se corre con el programa asigna16S.pl (ver anexos I) para lograr otro archivo con un formato mas legible que organiza todos los datos por el identificador del Genbank para el posterior parser, blastparser.pl (ver anexos II). Mediante este ultimo scrip, los datos taxonómicos son introducidos en la base de datos. El comando utilizado con el scrip asigna16s.pl es el siguiente:

$ perl asigna16S.pl fichero.fas.blastn > besthit.txt

El asigna16S.pl tiene dos formatos para correrlo, uno de ellos se basa en el

mejor hit por entrada, y el otro la mejor media en cada nivel taxonómico. De esta forma da el siguiente tipo de formato, (tras correrlo con el modelo para lograr la mejor puntuación en base a la media de cada nivel taxonómico):

```
# Creado por asigna16S.pl, Sat Sep  1 15:59:01 2012; Blast file: ggaa.fas.blastn; Database: greengenes
# Method: bestaver; mindiff=0.02; numtaxhits=5; mintaxhits=3

AB000684        superkingdom    Bacteria        Bacteria        97.46
AB000684        phylum  Aquificae       Aquificae       97.46
AB000684        class   Aquificae       Aquificae
AB000684        order   Aquificales     Aquificales     97.46
AB000684        family  Aquificaceae    Aquificaceae    97.46
AB000684        genus   Hydrogenobacter Hydrogenobacter 97.46
AB000684        species Unresolved      Hydrogenobacter subterraneus    95.65

AB000697        superkingdom    Unresolved      Bacteria        100.00
AB000697        phylum  Unresolved      Thermotogae     91.74
AB000697        class   Unresolved      Thermotogae
AB000697        order   Unresolved      Thermotogales   91.74
AB000697        family  Unresolved      Nautiliaceae    100.00
AB000697        genus   Unresolved      Caminibacter    100.00
AB000697        species Unresolved

AB001042        superkingdom    Bacteria        Bacteria        94.03
AB001042        phylum  Actinobacteria  Actinobacteria  94.03
AB001042        class   Actinobacteria  Actinobacteria
AB001042        order   Actinomycetales Actinomycetales 94.03
AB001042        family  Unresolved      Actinosynnemataceae     93.53
AB001042        genus   Unresolved      Lentzea 93.53
AB001042        species Unresolved      Lentzea violacea        93.53

AB001043        superkingdom    Bacteria        Bacteria        97.40
AB001043        phylum  Proteobacteria  Proteobacteria  97.40
AB001043        class   Alphaproteobacteria     Alphaproteobacteria     97.40
AB001043        order   Unresolved      Rhodospirillales        97.40
AB001043        family  Unresolved      Rhodospirillaceae       97.40
AB001043        genus   Unresolved      Magnetospirillum        95.63
AB001043        species Unresolved      Magnetospirillum magnetotacticum        95.63
```

Imagen.10: Resultado del asigna16s.pl, con el método de mejor media.

El programa asigna16s.pl puede trabajar de dos formas: Con la mejor media de cada nivel taxonómico o con la mejor puntuación de cada nivel taxonómico. En este caso se ha elegido la mejor media, que trabaja de la siguiente forma:

-En la lista de datos obtenida mediante el blast se escogen los mejores valores desde el nivel de reino hasta el nivel de especie. En cada nivel se coge el conjunto de valores con mejor puntuación media, y se desciende al siguiente nivel en base a la media obtenida en el paso anterior. En otras palabras, después de determinar cuales son las mejores puntuaciones a nivel de reino (bacteria) se desciende al siguientes nivel, phylum, y se escoge el phylum con mejor puntuación media respecto a la secuencia que se está analizando. De esta forma se analizan todos los niveles hasta la especie. El problema de este método es que necesita un mínimo de hits para realizar el calculo en cada nivel taxonómico. Esto se traduce, en que si un grupo taxonómico en concreto es muy pequeño, nunca se llegaría a nivel de especies, y se quedaría sin resolver la taxonomía. De la misma forma si la media no es muy buena en ninguno de los casos tampoco asigna ningún valor taxonómico a la misma.

-En el caso de trabajar con la mejor puntuación de cada nivel, solo se escoge el caso mas parecido a la secuencia que se está analizando. Esto en la mayoría de los casos dan valores buenos a nivel de genero y especie. El problema, es que la secuencia que se está analizando puede ser de una especie cercana, pero no la especie a la que estamos asignando. Este problema, aunque no es muy grabe, se va solucionando a medida que las bases de datos eferentes a la taxonomía de las especies conocidas, como es el caso de greengenes, vayan corrigiendo las filogenias en cada caso.

```
# Creado por asigna16S.pl, Thu Jul  5 16:55:38 2012; Blast file: ggaa.fas.blastn; Database: greengenes
# Method: besthit

AB000684        superkingdom    Bacteria        Bacteria        97.46
AB000684        phylum  Aquificae       Aquificae       97.46
AB000684        class   Aquificae       Aquificae
AB000684        order   Aquificales     Aquificales     97.46
AB000684        family  Aquificaceae    Aquificaceae    97.46
AB000684        genus   Hydrogenobacter Hydrogenobacter 97.46

AB000697        superkingdom    Bacteria        Bacteria        100.00
AB000697        phylum  Thermotogae     Thermotogae     91.74
AB000697        class   Thermotogae     Thermotogae
AB000697        order   Thermotogales   Thermotogales   91.74
AB000697        family  Thermotogaceae  Thermotogaceae  85.20
AB000697        genus   Thermotoga      Thermotoga      85.20

AB001042        superkingdom    Bacteria        Bacteria        94.03
AB001042        phylum  Actinobacteria  Actinobacteria  94.03
AB001042        class   Actinobacteria  Actinobacteria
AB001042        order   Actinomycetales Actinomycetales 94.03
AB001042        family  Unresolved      Actinosynnemataceae     93.53
AB001042        genus   Unresolved      Lentzea 93.53

AB001043        superkingdom    Bacteria        Bacteria        97.40
AB001043        phylum  Proteobacteria  Proteobacteria  97.40
AB001043        class   Alphaproteobacteria     Alphaproteobacteria     97.40
AB001043        order   Unresolved      Rhodospirillales        97.40
AB001043        family  Unresolved      Rhodospirillaceae       97.40
AB001043        genus   Unresolved      Magnetospirillum        95.63
```

Imagen.11: Ejemplo del resultado de del asigna16s.pl mediante el método de la mejor puntuación por entrada.

Por estas razones se decidió trabajar con el método de la mejor puntuación por hit. Aunque en algunos casos la identidad fiable se pierden a niveles de familia, ya que el mínimo de identidad se ha puesto a un 97% en especie, un 94% en genero y un 90% en el resto de niveles taxonómicos. Para ello, el programa blastparser.pl (ver anexos) realiza este filtro. En caso que para una secuencia el porcentaje de identidad sea inferior al marcado para cada nivel, no será guardado en la tabla de 'especies' en la columna de 'taxon'. De esta forma solo quedará guardada la entrada hasta el nivel taxonómico que resulte fiable.

En el caso de la tabla de especies la columna de 'taxon' se rellena con el siguiente formato (similar a la que encontramos en el archivo de greengenes.tax):

Bacteria(superkingdom);Aquificae(phylum);Aquificae(class);Aquificales(order);Aquificaceae(family);Hidrogenobacter(genus);Hidrogenobacter subterraneus(species)

De esta forma queda guardada la taxonomía para cada cluster que se haya definido mediante el CD-HIT-EST. Y en caso de querer llamar a un taxón en concreto resulta mas fácil y legible para el usuario.

Por otro lado el blastparser.pl (ver anexos II) carga todos los datos referentes al blast en la tabla de 'taxones'. En este caso guarda para cada identificador todos los valores para cada nivel taxonómico: mejor media, mejor hit y la puntuación para cada asignación. De esta forma siempre se puede ir a analizar todos los datos de cada entrada en caso de querer hacerlo.

Después de haber cargado todas las tablas de la base de datos, se pueden realizar nos análisis relacionales que se quiera, desde tipos de especies, secuencias, muestras, ambientes, … De esta forma con los datos que existen en cada momento se puede lograr valores estadísticos interesantes que reflejarán el conocimiento de la

distribución taxonómica del momento. Pero para ello, siempre hay que hacer una actualización de la base de datos cada cierto tiempo. Por ello, también se han preparado los programas para la actualización de la base de datos.

+Actualización de la base de datos:

Una vez cargada la base de datos si se quiere actualizar, los programas están preparados con un bucle para que solo se introduzcan las entradas nuevas. De esta manera no hace falta borrar ninguna entrada y las actualizaciones se realizan mas rápido. Lo programas utilizados se llaman prácticamente igual, solamente que llevan el prefijo ac- delante, para poder identificarlos. Cada programa de estos lleva a cabo un trabajo similar a los descritos con anterioridad en la carga de la base de datos por primera vez. Pero en este caso, lo programas lo que hacen es comprobar primero la existencia de datos en la base de datos. Posteriormente, en caso de que exista la misma entrada se salta sin hacer nada a la siguiente, en caso de que exista datos relacionados con la entrada (por ejemplo; tipo de muestra, cluster, …), se le relaciona la secuencia con dichos datos. Y por ultimo, si no existe ningún tipo de relación entre la entrada y los datos ya existentes, crea todo de nuevo. Para entenderlo mejor explicaremos todo el proceso programa por programa.

El caso del primer parser, acmicroparser.pl (ver anexos II), antes de introducir los datos en la base de datos se comprueba la existencia de dicha entrada en la base de datos. En caso de ser una nueva entrada la introducirá sin ningún problema, en caso de ser una entrada existente salta de bucle y pasa a buscar la siguiente entrada. El resto del programa es igual que el parser que se utiliza para rellenar la base de datos por primera vez.

De la misma forma solo se crean los archivos de fastas con las secuencias (con el nombre de outfasta2) y el archivo de los títulos y los autores para corregir de las nuevas entradas que se extraen con el parser.

A la hora de asignar el numero de muestra de las nuevas entradas se realizarán de la misma forma que las asignamos la primera vez. Es decir, si tienen el mismo titulo y los mismos autores se considerará la misma muestra. Para ello el proceso de corrección de títulos y autores se dará de la misma forma que en la primera carga de la base de datos. Solamente, que en este caso solo se realizaran las correcciones de las nuevas entradas que ha sido creada mediante el acmicroparser.pl . En esta ocasión, una vez realizada las correcciones pertinentes, el programa que introduce el numero de muestra en la base de datos tiene que abrir el archivo donde se quedo guardada la ultima numeración referente a las muestras. A partir de dicho numero empieza a contar otra vez para introducir el numero de muestra para las nuevas entradas.

Por otro lado también se tiene que comprobar que las las nuevas entradas no pertenezcan a una entra ya existente. Por ello, en el programa de actualización de la tabla de la base de datos, antes de asignar un nuevo numero de muestra busca en la base de datos el titulo y los autores de las nuevas entradas. En caso que ya exista se

le asigna a la nueva entrada el numero de muestra existente en la base de datos, que hace referencia al trabajo publicado. Por otro lado, en caso de ser una entrada que no está en la base de datos, se le asigna un nuevo numero de muestra. Esto lo lleva a cabo el programa acmuestratabla.pl (ver anexos II), con el cual se llena la tabla 'muestra' de las nuevas entradas.

El programa que rellena la tabla de 'ambiente', acmetadata_ambiente.pl (ver anexos II) funciona igual que el de acmuestratabla.pl. En caso de existir el trabajo de la nueva entrada en la base de datos, se le asigna el mismo identificador de ambiente. En caso de ser un trabajo totalmente nuevo, se le asigna un nuevo numero de ambiente.

Por otro lado, a la hora de actualizar las tablas que hacen referencia a las secuencias, especies y clasificación taxonómica, se trabajará de forma un poco diferente. Ya que por un lado las secuencias que hagan referencia a una especie ya descrita, simplemente habría que asignarle los mismos datos filogenético, y por otro lado,las especies nuevas habría que realizar el mismo proceso que se llevo a cabo en la primera carga de la base de datos.

Para realizar la clusterización de las secuencias nuevas se lleva a cabo con las secuencias que fueron guardadas en el archivo de la ultima carga de la base de datos, en el que ya solo están las secuencias fasta de las secuencias representantes de cada cluster. Para ello mediante un cat (orden UNIX) se concatenan los dos archivos, y el archivo resultante se vuelve a pasar por el CD-HIT-EST:

-Orden para concatenar los dos archivos:

$ cat outfasta* > fasta3

-La orden para que el CD-HIT-EST realice el análisis:

$ cd-hit-est -i fasta3 -o outfasta3 -c 0.97 -M 4000 -T 2 -aL 0.8 -aS 0.4 -l 200 -r 1

Con este sistema las nuevas secuencias se clusterizan y con otro programa actualizacion_clusteres.pl (ver anexos II) se le asignan a los que ya existen o crean sus propios clusteres. En otras palabras, las secuencias que estén dentro de un cluster ya existente simplemente se les asignarán a sus respectivos, y en los casos en el que las secuencias no están dentro de los ya descritos, se crean sus consiguientes clusteres. En este caso las secuencias que hayan creado un nuevo cluster se añadirán a un nuevo archivo de fastas, sobre el que se realizara el blast. Posteriormente el resto de programas para la la asignación de las especies y taxones, son los mismos, solamente que solo se trabaja con las nuevas secuencias que se hayan logrado.

El blast que posteriormente se va a realizar solo se llevara a cabo con las nuevas secuencias. De esta manera se optimiza mucho mas el tiempo de análisis. Posteriormente se sigue insertando los taxones en la base de datos mediante el

programa blastparser.pl (ver anexos II). Con el cual se guardarán todo los datos referentes a la taxonomía de cada secuencia nueva, para el posible análisis posterior de las mismas en caso de querer hacerlo.

Al finalizar la actualización de la base de datos quedarán guardados otra vez en archivos independientes, datos que nos serán útiles para las siguientes actualizaciones: el numero de muestra, el identificador de ambiente y el archivo de fastas en el que solo quedan las secuencias clusterizadas. De esta forma a la hora de actualizar las próximas veces, no habrá ningún problema de asignación.

Todos estos scrips, tanto los de cargar la base de datos como los de actualización están ordenados mediante un programa que coordina a todos, coordinador.pl (ver anexos II). Mediante este programa se organiza todos estos programas ya mencionados como todos los comandos Unix que se tienen que llevar a cabo a lo largo de todo el proceso de carga de la base de datos, así como la de actualización.

Este scrip funciona como un menú, con el cual el autor podrá elegir que tipo de paquete de programas quiere utilizar:

- Los programas que están creados para cargar la base de datos “de novo”.
- Los programas que están creados para actualizar la base de datos.

La diferencia entre los programas de un paquete y del otro, es que los scrips de actualización primero buscan en la base de datos la existencia de dicha entrada, antes de introducir nada. Por otro lado ciertos nombre son modificados para evitar problemas (sobretodo los referentes a los archivos con los que hay que trabajar). Y en este caso, a la hora de realizar las actualizaciones, hay que trabajar de forma coordinada junto con los archivos existentes, por ejemplo en el caso de hacer el CD-HIT-EST. En este caso se concatenan el fasta logrado de las nuevas entradas y el archivo de fasta devuelto por el CD-HIT-EST cuando se cargó la base de datos por primera vez.

De todas formas antes de correr este programa es necesario tener todas los programas externos que se necesitan para llevar a cabo todos los procesos. Los programas que se necesitan serian los siguientes:

- El blast: está para descargar desde la terminal como blastall con el siguiente comando:

$ sudo apt-get install blastall

- El CD-HIT: se tendría que descargar de su pagina (ver anexos I).
- El modulo LevenshteinXS: se tiene que descargar como modulo de cpan (ver anexos I).

Es importante tener estos programas descargados e instalados. Ya que de lo contrario el paquete daría problemas de programación y la carga de la base de datos quedaría anulada.

Por otro lado también tendría que estar instalado el scrip taxbuild_NOT_EUK.pm (ver anexos I) en el directorio de los módulos de perl. Ya que sin este scrip no se podría dar formato a los archivo del greengenes y no se crearía la base de datos contra la que se realizaría el blast.

En el momento que el scrip de coordinador.pl funcione y cargue por completo la base de datos, o la actualice, podremos empezar a analizar los datos que se hayan guardado.

+ Análisis de la información de la base de datos:

Una vez cargada por completo la base de datos, podemos empezar a analizar los datos recogidos en ella. Estos análisis pueden ir desde del tipo bibliográfico; como cuantos artículos se han publicado, o cuantas muestras se han analizado, o incluso como ha ido creciendo en numero de secuencias y de especies en la base de datos del Genbank, etc ; hasta crear redes de interacción entre bacterias en diferentes ecosistemas (los que se hayan descrito en la base de datos).

Para ello se han realizado varios scrips con los que han logrado diversa información:

-Evolución de las entradas y secuencias del Genbank a lo largo de los años:

Mediante el scrip esp_seq.pl seguido del programa conteofecha.pl (ver anexos III) podemos analizar como a lo largo del tiempo como ha ido evolucionando las entradas de secuencias y especies en la base de datos del Genbank. Este ultimo programa nos devuelve un archivo que posteriormente lo tratamos con el scrip en R de fecha_seq.R (ver anexos III), con el cual se logrará un gráfico en el que quedará reflejada los cambios en la cantidad de entradas que se han introducido en el Genbank.

-Cuantos muestras diferentes existen:

Mediante el scrip idmuestra.pl (ver anexos II) podemos ver realmente cuantas muestras diferentes se han analizado en busca de especies nuevas. Como se han asignado el numero de muestras en base a tener el mismo título y los mismo autores, el numero de muestra total se puede conseguir con un simple contador de lineas, con el siguiente comando:

$ wc -l unititautfin.txt

-Cuantos medios de extracción existen:

Con el scrip iso_sour.pl (ver anexos III) se extraen todos los tipos de medio de extracción que existen en l base de datos. Posteriormente con el archivo que crea aparecen todas las entradas existentes referentes a los medios. En caso de querer saber cuanto medios diferentes existes simplemente realizando un sort y posteriormente un uniq (comandos UNIX) sobre el archivo resultante obtendríamos el resultado. En otras palabras, simplemente marcando estos codigos depues de correr el programa obtendriamos la información que solicitamos:

$ wc -l isolation.txt

Se lograría el total de entradas que tienen asignado un medio de extracción.

$ sort isolation.txt > sortiso.txt

$ uniq sortiso.txt > uniqiso.txt

$ wc -l uniqiso.txt

Se lograría el todos los tipos de medios de extracción diferentes que existen.

Además este último archivo que se logra se podría utilizar para la creación de un clasificador en base a los medios que en él se mencionan. Este clasificador tendría que hacer frente a una gran cantidad de sinónimos y palabras clave que cada autor utiliza a su manera para referirse al mismo medio.

-Redes de interacción biológica:

Para dicho análisis es necesario además de tener la base de datos con todos los campos llenos, tener la base de datos actualizada. Ya que los análisis de interacción entre especies serán mas completos cuanto mas datos existan.

Este apartado tiene numerosos problemas, ya que en muchas entradas del Genbank, no se han descrito los medios en los que las muestras fueron obtenidas, ni dan información al respecto. Por lo que se ha tomado como referencia cada articulo como una muestra. En el caso concreto de las muestras en las que se especifica que tipo de medio en el que fue extraída, se tendrá en cuenta dicho medio.

Simplemente realizando una búsqueda por muestras y por medios se logran sacar cuantas especies se encuentran en cada medio. Posteriormente se analiza la interacción entre especies en base a la presencia-ausencia en diferentes medios. Este análisis de interacción se puede llevar a cabo con un estadístico de fisher. Para ello se podría utilizar un programa en R que analice los casos de presencia y ausencia de

diferentes especies en diferentes medios. En otras palabras habría que tener en cuenta en el archivo todas estas características:

Especie A, especie B, Nº de veces que aparece A, Nº de veces que aparece B y el Nº de veces que aparecen las dos juntas.

Estos datos serán extraídos de la base de datos mediante un scrip de perl, y posteriormente, con estos datos se podría crear un scrip en R para realizar el estadístico. Y posteriormente las entradas que lo pasen el umbral del 97% en el test de fisher se creará un nexo de unión entre dichas especies en un archivo para posteriormente poderlo verlo como resultado final en el cytoscape. Pero esta parte no se ha podido llevar a cabo. Por lo tanto sería el siguiente paso a llevar a cabo.

Resultados:

En lo que se refiere a la cantidad de secuencias existentes en la base de datos del Genbank, cabe destacar la reducción del total de secuencias que se han introducido durante el año 2011. Pese a ello la cantidad de datos referentes al 16S se está aproximando a 3'5 millones de entradas totales, de hecho actualmente existen 3.349.676 entradas, y sigue el aumento de su numero.

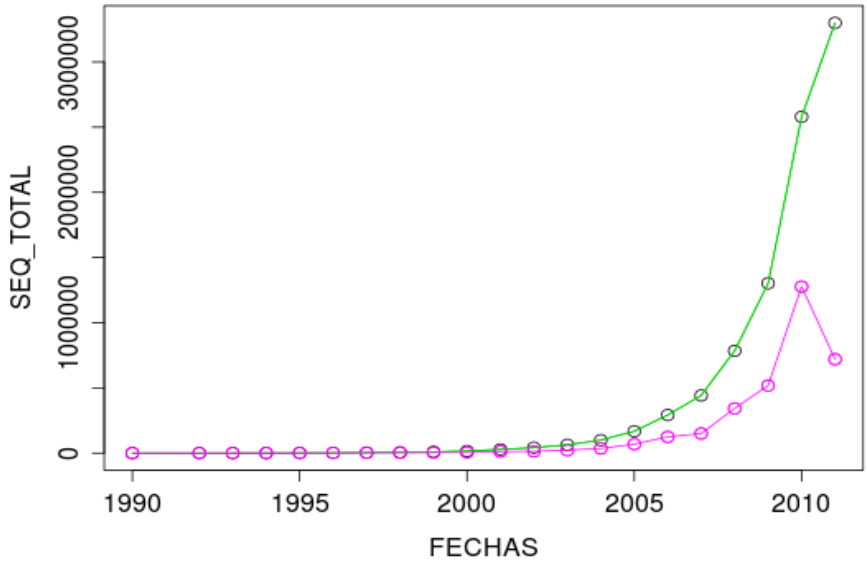


Imagen.12: Gráfico de la evolución de las secuencias en la base de datos del Genbank (1990-2011). En verde aparecen el numero de entradas totales y en rosa el numero de entradas ingresadas cada año.

En este gráfico podemos observar que cada año que pasa el numero de secuencias existentes en el Genbank va en aumento. Aunque el numero de secuencias introducidas en el año 2011 descendió considerablemente. Por otro lado, también se han logrado el numero de clusteres que se han identificado para cada fecha.

Asumiendo, que cada cluster se refiere a una especie diferente podríamos ver la evolución del descubrimiento de especies nuevas de año en año. Esto no quiere decir que todas las especies tengan que estar descritas. Ya que el numero de especies descritas en la base de datos de greengenes solo son 10.000.

En lo que se refiere a las muestras, se han logrado 11.661 muestras diferentes, en base a la diferenciación entre títulos y autores llevada a cabo por el scrip idmuestra.pl (ver anexos II). Esto nos demuestra que todas las entradas existentes en el Genbank, referentes a la descripción del gen 16S en las bacterias, se han llevado a cabo en 11.661 trabajos diferentes. Pero eso no quiere decir en todos esos trabajos se hayan realizado en el mismo medio ecológico. De hecho, existen trabajos que tienes mas de un medio de extracción.

Por otro lado cabe destacar que en todas estas entradas existen 11.071 medios de extracción diferentes. También tenemos que tener en cuenta que una parte de las entradas no tienen ningún dato referente al medio en el que fue extraído. Para hacerlo mas visual:

En la base de datos existen 3.349.676 entradas diferentes, de las cuales, 3.031.987 tienen algún tipo de identificador referente al medio en el que fueron extraídas. Estas etiquetas referentes a los medios son bastante pobres y por ello se ha decidido aprovechar a coger también los datos referentes a la nación, hospedador y/o coordenadas que se pueden extraer también. Con lo que se puede hacer un perfilamiento mas concreto para cada caso.

Por otra parte, se ha visto que a la hora de hacer una clasificación mediante el medio de extracción seria muy complicado. La razón de esto es que no están descritas de una forma estandarizada, en otras palabras, existen medios que podrían estar relacionados, ya que hacen referencia al mismo entorno, pero cada autor le da diferente nombre. Por ejemplo: sea, sea water, marine water, ocean, coastal water, … En estos casos, es prácticamente imposible realizar una clasificación con un buen criterio.

Para solucionar este problema se podría recurrir a la técnica de text-mining y clasificar los elementos por las palabras que tenga. De la misma forma habría que realizar una clasificación en base a un diccionario de sinónimos para dichas palabras. Al menos de la forma que está creada la base de datos solo habria que extraer los campos de metadatos y isolation_source y en base a ellos hacer un clasificador de medios.

Por otro lado, en lo que es referente a la sección de las secuencias y especies, cabe destacar la existencia de 592.548 clusteres, que se podria traducir cada uno como una especie. Aunque cabe destacar que multiples especies tienen difentees cepas en las el 16S lo tiene modificado. Aun así, este dato seria interesante, ya que en el día de hoy se tienen descritas unas 10.000 especies de bacterias, y por otro lado se tiene estimada que el numero total de especies de bacterias existentes en el planeta ronda entre 2-5 millones. De todas formas antes de hacer esta afirmación

habría que realizar mas análisis respecto a las muestras que se hayan extraído. Ya que en este caso estamos asignando especies solamente con el gen 16S.

Respecto a las tablas taxonómicas que nos devuelve el asigna16s.pl podemos ver los diferentes datos logrados para cada asignación taxonómica (imagen). Esta seria la clasificación obtenida para todas las secuencias que se han analizado.

Esta clasificación nos da los siguientes datos taxonómicos:

- A nivel de reino se pueden clasificar en Bacterias (539.369), Archaeas (27.520) y secuencias sin clasificar (23.832).
- Se clasifican en 30 phylums todas las entradas.
- A nivel de clase, se han descrito 52 clases diferentes.
- Se han encontrado 111 ordenes.
- Respecto a las familias se han descrito 272 diferentes.
- Se han descrito 1414 géneros diferentes.
- A nivel de especies se han clasificado en 6001 especies diferentes.

El resto de datos referentes a este programa están guardados en el archivo adjunto de TAXONOMIA.xls (archivo anexo). En este archivo podemos ver todos los datos para cada nivel taxonómico que se le haya asignado a cada caso, así como la cantidad de secuencias que se han encontrado para cada caso.

Cabe destacar que varias especies se les ha asignado múltiples secuencias. Como es el caso de Clostridium sp. que se le han asignado un total de 11.089 secuencias. Estas asignaciones tan numerosas pueden significar dos cosas: o la existencia de varias subespecies o la clasificación errónea por diversos factores que posteriormente discutiremos.

Discusión/Conclusiones:

El hecho que observemos 592.548 clusteres que se podrían interpretar como diferentes especies es un dato muy interesante. Aunque partiendo de la base que solo se tienen descritas 10.000 especies bacterianas en el greengenes, la asignación que hemos realizado de mejor puntuación por nivel taxonómico no nos va a decir grandes cosas. De hecho en los resultados del blast ya se ve que varias especies tienen múltiples secuencias. Esto se puede interpretar de varias maneras:

El hecho que varias secuencias se hayan asignado a la misma especie puede significar que existen diferentes subespecies para esa especie y además da la

casualidad que tienen el gen 16S con una similitud inferior al 97%.

Que varias secuencias se hayan clasificado en la misma especie puede haber sido por lo siguiente; cuando se utilizó el programa asigna16s.pl se analizó mediante el mejor hit para cada nivel taxonómico. Esto puede provocar que especies cercanas a la que se le han asignado las secuencias, pero que no estén descritas, se sean directamente introducidas en la misma especie. Este error se irá corrigiendo a medida que se vayan describiendo nuevas especies. De momento, solo podemos asignar estas secuencias al mismo taxón, ya que no tenemos mas información.

Por otro lado se podría realizar otra clasificación mediante la mejor media por nivel taxonómico mediante el programa asigna16s.pl. Esto nos dejaría cada secuencia en el grupo taxonómico que mas se le parece. Aunque los valores serian muy malos, sobretodo para la hora de asignar especies o géneros. Por ello el gran problema es que la falta de taxones hace imposible una asignación fiable en el día de hoy. Habría que esperar a que la clasificación taxonómica de las especies se vaya completando cada vez mas. Ya que intentar clasificar mas de medio millón de secuencias con una diferencia superior al 3%, que es el limite propuesto de forma mas universal para asignar la taxonomía en las bacterias, en base a que solo existen 10.000 especies descritas, no es un método acertado.

De todas formas querer asignar un taxón a cada muestra, en base únicamente al gen 16S, no se puede considerar un método muy fiable. Ya que para diferenciar adecuadamente las bacterias es necesario la secuenciación de todo su genoma y la posterior comparación del mismo entre todas ellas. Ya que muchas especies pueden coincidir en la secuencia de varios genes, pero el conjunto de sus genomas sean diferentes.

Por ello, este método podría considerarse como una aproximación taxonómica de las bacterias que se estén analizando. Ya que clasificar todas las bacterias en base a un solo gen no se puede considerar como algo de un peso lo suficientemente importante como para no poder plantearse otras opciones mas fiables. Pero el hecho que ya existan varias clasificaciones en base a la secuencia de este gen en diferentes especies, facilita el trabajo. Aunque no se puede olvidar en ningún momento que es simplemente un boceto referente a la clasificación microbiana.

De la misma forma, en caso de querer relacionar diversas especies mediante los medios en las que fueron recogidas van a dar varios problemas de relación. Ya que la asignación taxonómica no es del todo fiable y la relación de muestras y ambientes es muy difícil de llevar a cabo. Ya que la información referente a los medios de extracción precisaría de text-mining y de unificación de criterios, para que sea eficaz. Puesto que al no haber una norma que regule la entrada de dicha información en la base de datos del Genbank provoca un gran caos a la hora de asignar a cada muestra un medio concreto. Pero que se le podría dar una solución con la utilización de otros métodos, que en este trabajo no se han desarrollado.

Como conclusión final podríamos decir que la base de datos cumple todos los

requisitos para analizar la información referente a los medios y muestras así como la información relacionada con los taxones que en ella aparecen. Pero quedaría solucionar los problemas de la asignación de los medios de extracción y la mejora a la hora de asignar taxones de forma mas correcta. A medida que esos dos problemas se vayan solucionando el posterior análisis de la información de la base de datos será mucho mas fiable y mas correcta.

Bibliografía:

[1]: Metatranscriptomic Approach to Analyze the Functional Human Gut Microbiota. Marıa Jose Gosalbes, Ana Durban, Miguel Pignatelli, Juan Jose Abellan, Nuria Jimenez-Hernandez, Ana Elena Perez-Cobas, Amparo Latorre, Andres Moya. (2011), PLoS ONE 6(3): e17447. doi:10.1371/journal.pone.0017447.

[2]: The ‘‘Most Wanted’’ Taxa from the Human Microbiome for Whole Genome Sequencing. Anthony A. Fodor, Todd Z. DeSantis, Kristine M. Wylie, Jonathan H. Badger, Yuzhen Ye, Theresa Hepburn, Ping Hu, Erica Sodergren, Konstantinos Liolios, Heather Huot-Creasy, Bruce W. Birren, Ashlee M. Earl. (2012), PLoS ONE 7(7): e41294. doi:10.1371/journal.pone.0041294.

[3]: Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Nicola Segata, Susan Kinder Haake, Peter Mannon, Katherine P Lemon, Levi Waldron, Dirk Gevers, Curtis Huttenhower and Jacques Izard. (2012), Genome Biology 13:R42.

[4]: Activity of abundant and rare bacteria in a coastal ocean. Barbara J. Campbella,1, Liying Yua, John F. Heidelbergb, and David L. Kirchmana. (2011), PNAS, vol. 108, no. 31, 12776–12781.

[5]: Environmental distribution of prokaryotic taxa. Javier Tamames, Juan José Abellán, Miguel Pignatelli, Antonio Camacho, Andrés Moya. (2010), BMC Microbiology 2010, 10:85.

[6]: Diversity of 16S rRNA Genes within Individual Prokaryotic Genomes. Anna Y. Pei, William E. Oberdorf, Carlos W. Nossa, Ankush Agarwal, Pooja Chokshi, Erika A. Gerz, Zhida Jin, Peng Lee, Liying Yang, Michael Poles, Stuart M. Brown, Steven Sotero, Todd DeSantis, Eoin Brodie, Karen Nelson, and Zhiheng Pei. (2010), Applied and Environmental Microbiology, Vol. 76, No. 12.

[7]: Micro-Mar: a database for dynamic representation of marine microbial biodiversity. Ravindra Pushker, Giuseppe D'Auria, Jose Carlos Alba-Casado and Francisco Rodríguez-Valera. (2005) BMC Bioinformatics, 6:222 doi:10.1186/1471-2105-6-222.

[8]: Phylometrics: a pipeline for inferring phylogenetic trees from a sequence relationship network perspective. Samuel A Smits, Cleber C Ouverney. (2010) BMC Bioinformatics, 11(Suppl 6):S18.

[9]: VITCOMIC: visualization tool for taxonomic compositions of microbial communities based on 16S rRNA gene sequences. Hiroshi Mori, Fumito Maruyama and Ken Kurokawa. (2010) BMC Bioinformatics, 11:332.

[9]: TaxMan: a taxonomic database manager. Martin Jones and Mark Blaxter. (2012) BMC Bioinformatics, 7:536 doi:10.1186/1471-2105-7-536.

Anexos con programas y material complementario:

+ANEXO I (Programas y módulos ya existentes):

-CD-HIT-EST:

- cd-hit-est -i fasta.fa -o outfasta -l 200 -c 0.97 -M 4000 -aL 0.8 -aS 0.8 -r 1
- -i :archivo de entrada
- -o :archivo de salida
- -l :longitud mínima de secuencias
- -c :threshold
- -M :uso de memoria máxima
- -aL :coverage de la cadena larga
- -aS :coverage de la cadena corta
- -r para comprobar las cadenas en +/+ y +/-

-BLAST:

– blastall -p blastn -i outfasta -d greengenes -o file.out -m 8 -e 1e-03 -D 200 -a 2

– -p : el programa blastn para nucleotidos

– -i : el archivo de entrada (un fasta)

– -d :la base de datos que se va a utilizar para lanzar las secuencias

– -o : el archivo de salida.

– -m : formato del archivo de salida

– -e : el evalue minimo

- -D : la cantidad de salidas para cada entrada
- -a : la cantidad de cpu a utilizar.

-Text-LevenshteinXS:

-http://search.cpan.org/dist/Text-LevenshteinXS/LevenshteinXS.pm

- Para instalar un módulo cpan:
- $ sudo perl -MCPAN -e shell
- entrar en la shell, e instalar el modulo Text-LevenshteinXS :
- $ install Text-LevenshteinXS

-asigna16S.pl:

#scrip creado por Javier Tamames

-mtax.pl:

#scrip creado por Javier Tamames

--taxbuild_NOT_EUK.pm:

#scrip creado por Javier Tamames

+ANEXO II (Programas propios):

-microparser.pl
-idmuestra.pl
-muestratabla.pl
-metada_ambiente.pl
-cdhitparser.pl
-conteofasta.pl
-blastn.pl
-blastparser.pl
-coordinador.pl
-acmicroparser.pl
-acmetadata_ambiente.pl
-actualizacion_clusteres.pl
-acconteofasta.pl
-bdmicro.sql

+ANEXO III (Programas para el análisis de la base de datos):

-esp_seq.pl
-conteofasta.pl
-fecha_seq.R