

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



- Mikel Aguirre Rodrigo

Esquema

- Introducción
- La obtención de datos
- La base de datos
- Como se carga la base de datos
- Actualización de la base de datos
- Análisis de la base de datos
- Conclusiones

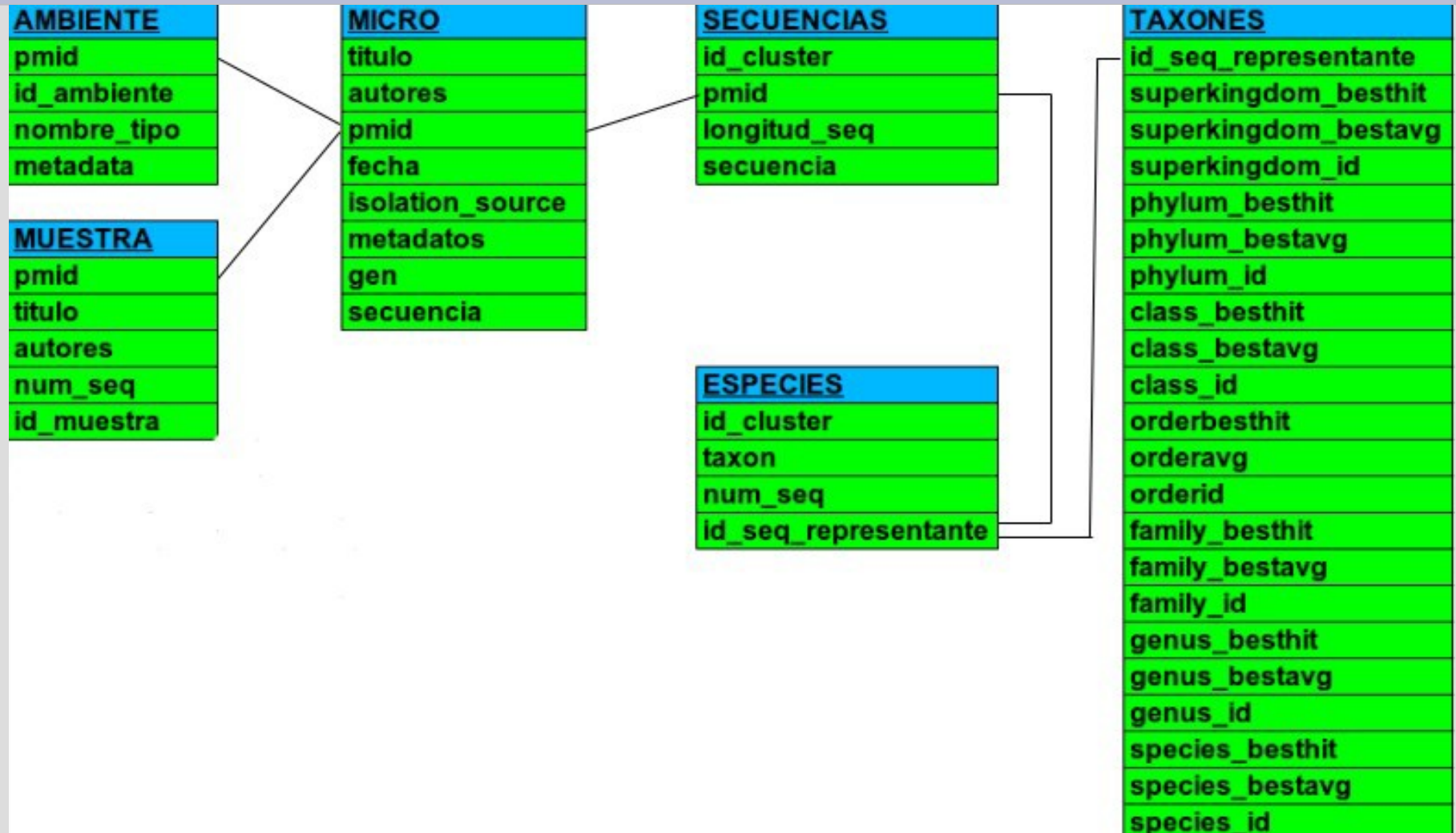
Introducción

- Desde que se se utilizan las técnicas de secuenciación masiva (metagenómica) se describen mejor las interacciones entre las diferentes especies de cada ecosistema.
- Para ello es necesario analizar cada caso extrayendo toda la información posible referente al medio y a la especie
- A medida que se realizan más metagenomas, la información que se puede extraer de cada medio será mas fiable a la hora de caracterizar ecosistemas

La obtención de los datos

- Los datos con los que va a trabajar son procedentes del Genbank.
- De esta base de datos se van a extraer solo las secuencias referentes al gen 16S de las bacterias.
- De las entradas referentes al 16S, se va a realizar una parser sobre los campos con mayor interes para este trabajo.

La base de datos



Cargar la base de datos

- Bajar los archivos de la base de datos del Genbank.
- Llevar a cabo el parser sobre los archivos y guardarlos en la base de datos.
- A la hora de realizar el scrip hay que tener en cuenta numerosas diferencias entre las diversas entradas.

LOCUS E0748459 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: agagtttgatymtggtcag, rev_name: 1513R, rev_seq: tacggytaccttggtacgactt"

rRNA <1..>1360

/product="16S ribosomal RNA"

ORIGIN

1 gatgaacgct agctacaggc ttaacacatg caagtcgagg ggaaacgacg gcgggggttc

61 ggccttgccg ggcgtcgacc ggcggatggg tgagtaacgc gtatccaacc tgcccctgtc

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

source Location/Qualifiers

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 ccccgttaaca

121 taacgatgtg gcatcacatt gttattatag cttcggcgct tcttgatggg tgtgcggtg

181 attagatagt tggcggggta acggcccacc aagtctacga tcagtagctg atgtgagagc

241 atgatcagcc acacgggcac tgagacacgg gcccgactcc tacgggaggc agcagtaagg

Una vez cargada la primera tabla

- Datos referentes al medio de extracción de la muestra:
 - Muestra (mediante el título y los autores)
 - Medio de extracción
- Datos referentes a la especie y la secuencia:
 - Secuencias
 - Taxonomía

Para realizar la asignación de la muestra

- Muestra = mismo título y autores
- Hay que corregir los títulos y autores
 - Modulo cpan: LevenshteinXS

Analysis of the archaeal sub-seafloor community at Sulu Seamount on the Izu Bonin Arc
Analysis of the bacterial communities associated with subtropical white syndrome of the coral *Turbinaria mesenterina* by oligonucleotide fingerprinting of ribosomal genes
Analysis of the bacterial communities associated with Subtropical White Syndrome of the coral *Turbinaria Mesenterina* by Oligonucleotide Fingerprinting of Ribosomal Genes
Analysis of the bacterial communities in continuous cotton fields of Xinjiang Province using 16/18S rDNA PCR-DGGE

Atam, S.I., Dube, S., Agarwal, N.K. and Singh, L.
Alavandi, S.V., Saravana Kumar, C., Dineshkumar, N., Kalaimani, N. and Poornima, M.
Alavandi, S.V., Saravana Kumar, C., Dinesh Kumar, N., Kalaimani, N. and Poornima, M.
Alavandi, S.V., Saravana Kumar, C., Dineshkumar, N., Poornima, M. and Kalaimani, N.
Alavi, M., Miller, T., Erlandson, K., Schneider, R. and Belas, R.

16S rRNA partial sequence of rumen archaea clone IVRI-RL-001 from buffalo (*Bubalus bubalis*)
16S rRNA partial sequence of rumen archaea clone IVRI-RM 002 from buffalo (*Bubalus bubalis*)
16S rRNA partial sequence of rumen archaea clone IVRI-RM 003 from buffalo (*Bubalus bubalis*)
16S rRNA partial sequence of rumen archaea clone IVRI-RM 004 from buffalo (*Bubalus bubalis*)
16S rRNA partial sequence of rumen archaea clone IVRI-RM 005 from buffalo (*Bubalus bubalis*)
16S rRNA partial sequence of rumen archaea clone IVRI-RM 006 from buffalo (*Bubalus bubalis*)
16s rRNA partial sequence of *Staphylococcus* sp., isolated from fish slime of *Leioqnathus bindus* collected from Ennore fish market

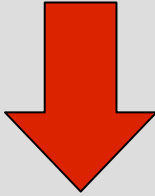
Medio de extracción

- Se extrae los siguientes campos de cada entrada (si las tiene):
 - Medio de extracción
 - Hospedador
 - Coordenadas
 - País

Asignación de las especies

- Las secuencias se guardan en un fasta y se realiza los siguientes pasos:
 - Eliminar las posibles secuencias con redundancias.
 - Realizar un blast con las secuencias representantes del resultado anterior.

Eliminación de secuencias redundantes

- Mediante el DC-HIT-EST:
 - Porcentaje de similitud: 97%
 - Coverage : 80%
- De 3.349.676 secuencias totales

- A 592.548 clusteres


Asignación de la taxonomía

- Un blast mediante el programa blastall :
 - E-value : 1e-03
 - Tendiendo como base de datos taxonómica el greengenes:
 - En este caso hay que bajarse el archivo en el que están todas las especies y secuencias relacionadas.
 - Dicho archivo hay que darle un formato especial mediante ciertos programas

Guardar y relacionar cada secuencia con su taxonomía

- A cada secuencia se le asigna una taxonomía en base a la mejor puntuación en cada nivel taxonómico.
- Se guarda todos los datos recogidos en el blast
- A cada secuencia se le asigna la taxonomía de la siguiente forma:
 - Especies: desde 97%
 - Géneros: desde 94%
 - Familia y resto de niveles taxonómicos: 90%

Actualización de la base de datos

- Este paquete de programas realizan los mismos pasos pero con un filtro:
 - Si ya existe en la base de datos, no lo introduce, y pasa a la siguiente entrada.
 - En el caso de las muestras:
 - Primero busca si pertenece a una muestra ya existente, y posteriormente asigna la entrada a dicha muestra
 - En caso de que no encuentre ninguna
- 
- Crea una muestra nueva

Actualización de la base de datos

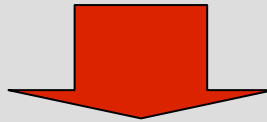
- Para asignar las especies de las nuevas entradas, el primer parser crea un fasta solo con las secuencias nuevas.
- El fasta de las secuencias nuevas se concatena junto con el fasta que devolvió el primer CD-HIT-EST con las secuencias sin redundancias
- Posteriormente se realiza otra vez el CD-HIT-EST sobre ese archivo con el fin de encontrar clusters nuevos

Actualización de la base de datos

- Con las secuencias que creen un cluster nuevo se realizará un blast y posteriormente se les asignará el taxón correspondiente a su tabla.
- Las secuencias que se hayan asignado a un cluster ya existente se les asignará su propia taxonomía.

Análisis de la base de datos

- El fin de todo este proceso de clasificación era poder clasificar los taxones en los medios descritos.



- Analizar las interacciones entre especies
 - Las estructura microbianas de ciertos medios
- También se puede sacar información bibliográfica de la base de datos: la evolución de la cantidad de secuencias y especies en la base de datos a lo largo de los años, las publicaciones de cada autor, ...

Análisis de la base de datos

- Para que la clasificación de especies en medios ecológicos sea fiable habría que realizar un análisis mediante text-mining:
 - Porque no existe un criterio a la hora de asignar el medio (agua marina, mar, océano, aguas de la costa, ...)
- Es recomendable tener la base de datos actualizada para que todos los errores que puedan existir referentes a la taxonomía se puedan ir corrigiendo.

Conclusiones

- Una vez cargada la base de datos se puede sacar múltiple información referente a la taxonomía y a los medios ecológicos
- Es un trabajo incompleto, y por lo tanto todavía quedaría trabajo que realizar:
 - Text-mining para los medios ecológicos
 - Análisis de interacción entre especies

Muchas gracias!