LCG-BioInfo

First version: 2022-09-19; Last update: 2022-11-09

EXAMEN 2 DE INTRODUCCIÓN A LA BIOINFORMÁTICA

Objetivo

El objetivo de éste examen es evaluar si el alumno entendio los conceptos básicos de la comparación de secuencias, el uso de scripts y awk.

Entregables

Los entregables del examen son:

a) Reporte de Análisis de datos en formato markdown.

b) En el servidor, un directorio llamado examen2 que tendrá todo los resultados.

Instrucciones

- Para cada pregunta, agrega una breve descripción.
- La respuesta agregala en un bloque de código [Si el bloque trae otros elementos además de código y comentarios, no será evaluado]. Si tienes scripts, deberás indicar en éste protocolo la forma de ejecutarse o lo que se corrió para obtener los resultados.
- Dentro del bloque o el script, no olvides poner comentarios que indiquen el algoritmo o lo que estas haciendo.

Criterio de Evaluación

Buenas prácticas

- Archivos y directorios.
  - Nombrado. Se evaluará que los archivos y directorios sean nombrados de acuerdo al contenido.
  - Nombres aceptables. Los nombres NO deben tener espacios ni caracteres especiales. Solo carcateres alfanumericos y - o _
- Organización. La estructura de directorios que se ha usado durante las clases deberá implementarse para éste exámen. Cada archivo deberá estar en su carpeta correspondiente.
- Código legible y reproducible

Importante. Puedes hacer uso de internet, checar tus apuntes. Pero, no puedes preguntar o copiar. Nuestro interés es conocer tus habilidades y lo que necesitamos reforzar.

Los puntos extra se aplican a la calificación final del examen, donde la calificación máxima es 10.

SECCION TEORICA [20%]

1. ¿Qué es y para que sirve un alineamiento de secuencias?
2. ¿Qué tipos de alineamientos existen?
3. ¿Qué tipo de alineamientos realiza la familia de programas de BLAST?
4. ¿Que requiere el programa blastp o blastn para poder ejecutarse (menciona al menos 2 opciones/argumentos)?
5. ¿Qué es un ortólogo y bidirectional best hit?

SECCION PRACTICA [60%]

Protocolo

Introducción

La función principal en los organismos microbianos es controlar la respuesta a los cambios ambientales, como el estado nutricional y varios estreses. Una idea importante que surge en la biología posgenómica es que la regulación transcripcional puede verse como una red compleja de interacciones entre diversos tipos de moléculas como proteínas, ADN y metabolitos. En este trabajo tratamos de evaluar la evolución de la estructura y plasticidad de la red reguladora transcripcional (TRN) a través de las especies a través de los regulones [Un regulón se define como el grupo de todos genes regulados por un factor de transcripción (TF).], mediante un análisis comparativo de su conservación.

Con el fin de entender la plasticidad de la Red de Regulación en bacterias, vamos a estudiar la conservación de la red de regulación transcripcional a partir de un organismo modelo *Escherichia coli* comparandola a través de todos los bacterias secuenciadas hasta el momento.

Para ello se obtuvieron los ortólogos bidireccionales (BDBH bidirectional best hit) entre E. coli y todas las bacterias. Usando el programa blastp con e.value <= 10-3 con un 95% de identidad [1].

Es importante hacer notar existen otros criterios adicionales para indicar que un gene es un ortólogo de otro, por ejemplo si conserva sus dominios funcionales, pero para este análisis solo tomaremos en cuenta la homología por secuencia.

Planteamiento del problema

Datos de entrada

- Genes con su información para cada bacteria. La columna 3 es el identificador de genbank GI del gene, en cada organismo E_coli_K12.list S_typhi.list S_typhimurium_LT2.list

Formato:

```
#1)feature-type 2)geneName      3)GI    4)locus 5)left..rightGenomePos
6)strand        7)externalDBsIDs        8)annotationsFunctionProd
CDS     thrL    16127995        b0001   190..255        F       ASAP:ABE-
0000006;ECOCYC:EG11277;GeneID:944742   function="leader; Amino acid
biosynthesis: Threonine" function="1.5.1.8 metabolism; building block
biosynthesis; amino acids; threonine" product="thr operon leader peptide"
CDS     thrA    16127996        b0002   337..2799       F
ASAP:8;ECOCYC:EG10998;GeneID:945803     function="enzyme; Amino acid
biosynthesis: Threonine" product="bifunctional aspartokinase I/homeserine
dehydrogenase I"
```

- Ortólogos entre pares de organismos. La columna 1 contiene el identificador GI del gene de E. coli, y la columna 2, es el identificador GI del otro organismo comparado. E_coli_K12.S_typhi.bdbh E_coli_K12.S_typhimurium_LT2.bdbh

```
#1)QueryID      2)TargetID      3)-     4)-     5)-     6)Evalue
7)%queryInAlignment     8)%targetInAlignment    9)Query
identities/alignment length     10)Target identities/alignment length
11)Query alignment init,alignment end   12)Target alignment
init,alignment end  13)QueryLength  14)TargetLength 15)Specify Identical
16127996        16758995        1       1538    3981    0.0     100
100     774/820 797/820 1,820   1,820   820     820     BI-DIRECTIONAL
16127997        16758996        1       597     1540    1e-169  99
100     288/308 297/308 1,308   1,308   310     309     BI-DIRECTIONAL
16127998        16758997        1       790     2039    0.0     100
100     399/428 411/428 1,428   1,428   428     428     BI-DIRECTIONAL
```

- Red de regulación de E. coli https://regulondb.ccg.unam.mx/menu/download/datasets/files/network_tf_gene.txt

Lista de Factores transcripcionales de E. coli

https://regulondb.ccg.unam.mx/menu/download/datasets/files/TFSet.txt

```
# (1) Transcription Factor (TF) identifier assigned by RegulonDB
# (2) TF Name
# (3) TF Synonyms List
# (4) Gene Coding for the TF
# (5) TF Active Conformations
# (6) TF Inactive Conformations

mostrando las columnas de interes..

(1)             (2)     (4)
ECK125328145    AaeR    aaeR
ECK125286586    AccB    accB
ECK120015994    AcrR    acrR
ECK120012595    Ada     ada
ECK120014170    AdiY    adiY
ECK120012515    AgaR    agaR
ECK125134683    AidB    aidB
ECK120030264    AlaS    alaS
ECK120015630    AllR    allR
ECK120015636    AllS    allS
ECK120012984    AlpA    alpA
ECK120011959    AlsR    alsR
```

Metodología

Para éste análisis de evaluar la plasticidad de la red de regulación, sólo tomaremos como ejemplo 2 organismos S_typhi y S_typhimurium, para fines didácticos.

Recuerda que en cada pregunta, aunque la respuesta pueda ser un si o no, debes demostrar como llegaste a esa conclusión, asi que debes indicar el código o instrucción que te llevo a esa respuesta.

1. **Revisión de los archivos de entrada**

Por cada organismo indica lo siguiente (puedes crear una tabla con los resultados)

Archivos *.list

a. Total de genes de cada organismo que vienen en el archivo *.list
b. ¿Todos los genes traen nombre?
c. ¿ Todos los genes traen GI y no se repite ?
d. ¿Vienen lineas en blanco o vienen comentarios en el archivo ?

Archivos *.bdbh

e. ¿Cuantos ortologos bidireccionales trae cada archivo.?

f. ¿Vienen lineas en blanco o vienen comentarios en el archivo ?

g. Si crees que es necesario checar algo más de los archivos puedes hacerlo. [extra 0.5 pto]

2. **Sobre los resultados del Blast.**

a. Por cada archivo *.bdbh que contiene los ortologos bidireccionales, extrae cuantos genes de E. coli tienen el mismo tamaño al del organismo target (columna 13,14)

b. ¿Cuántos genes de E. coli además de tener el mismo tamaño, el alineamiento cubrió el 100% tanto de E. coli como del organismo Target ? (columna 7,8 y 13,14)

c. ¿Cuántos genes de E. coli tienen el mismo tamaño del organismo target , el alineamiento cubre el 100% de la secuencia de ambos organismos, y las identities cubren todo el gene? [extra 1 pto]

3. **Sobre la red de regulación**

a. Escoge un TF de la red de regulación, y usando el nombre del gene de TF obten su GI del archivo .list, y verifica que ese TF se encuentre en los archivos *.bdbh. En general, el nombre del TF CRP, su gene es crp, o AraC seria araC. Hay algunas excepciones como IHF que es un regulador heterodimero o complejo donde los genes que lo codifican es ihfA y ihfB. Selecciona un TF donde sea fácil identificar su gene codificador.

b. Usando el TF seleccionado en el punto anterior, obten sus genes regulados, e indica cuántos se conservan en los otros organismos.

c. Usar el archivo de TFSet.txt, obten el gene que lo codifica (columna4) y busca el nombre del gene en el archivo E_coli_K12.list, solo quedate con los reguladores cuyo gene que lo codifica es solo 1. Con esa lista de genes reguladores, busca si existe un ortólogo en los otros organismos. [extra 1 punto]

d. Para cada regulador de la red, obten sus genes regulados y verifica cuantos de ellos tienen un ortólogo y saca el porcentaje de conservación. (El GI te servirá para buscarlos.) [extra 1 punto]

RESULTADOS Y CONCLUSIONES [20%]

1. Agrega la estructura del proyecto en el reporte.

2. El archivo README, que es el reporte, debe estar también en el servidor dentro de la carpeta del proyecto.

3. Interpretar los resultados según los resultados obtenidos. ¿Qué puedes decir de la red de E. coli y su plasticidad?

4. Conclusiones de lo aprendido. Puedes ayudarte de las siguientes preguntas ¿Que aprendiste durante la realización de la práctica? ¿qué piensas de unix y sus comandos avanzados?

BIBLIOGRAFÍA

Sólo en caso de haber utilizado alguna otra fuente.

1. Bacterial regulatory networks are extremely flexible in evolution https://academic.oup.com/nar/article/34/12/3434/2375626