Universidad Nacional Autónoma de México

Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas

iimas

Documentación Software para el Tratamiento de Genes

Laboratorio de Redes Biológicas

Fecha de Actualización: 21/07/2022

Requisitos Iniciales

Python:

- Python v3.8.x ó Python v3.9.x -> https://www.python.org/

R:

- R v4.1.3 ó R v4.2.0 -> https://www.r-project.org/

Paquetes de Python

Una vez instalado Python, se requiere instalar (desde cualquier terminal o símbolo del sistema) una serie de paquetes mediante el comando:

pip install <nombre_paquete>

Donde *<nombre_paquete>* será reemplazado por el nombre del paquete a instalar de acuerdo con la siguiente lista:

- jupyter
- tqdm
- pandas
- numpy
- re
- gzip
- shutil
- os
- io
- cufflinks
- GEOparse
- pysradb
- bioinfokit
- gtfparse
- matplotlib
- subprocess
- PyPDF2
- pycombat *
- logging
- scipy

- warnings

Algunos paquetes se instalan de forma automática durante la instalación de otros por lo que puede no ser necesario instalarlos de forma manual.

* Nota: si al instalar pycombat se genera un error porque intenta reinstalar pandas, numpy o alguna otra biblioteca, ignorar el error y ejecutar el comando *pip install combat* hasta que no se generen errores.

Paquetes de R

La instalación de los paquetes de R necesarios para el funcionamiento del código se encuentra especificada dentro del mismo por lo que no es necesario llevar a cabo instalación de forma manual.

Instalación de Software Especial

SRAToolkit: utilizado para la descarga de archivos sobre experimentos ARN-seq alojados en la base de datos biológica NCBI.

- sratoolkit v3.0.0

Instalación en Windows:

- Descargar el siguiente archivo MS Windows 64 bit architecture
- Descomprimirlo en la ruta que el usuario desee.

Instalación en Linux (Sistemas basados en Ubuntu):

- Dentro de la terminal colocarse en la ruta deseada para el almacenamiento del software.
- Ingresar el siguiente comando:

  wget --output-document sratoolkit.tar.gz https://ftp-trace.ncbi.nlm.nih.gov/sra/sdk/current/sratoolkit.current-ubuntu64.tar.gz
- Extraer el archivo con ayuda del comando:

  tar -vxzf sratoolkit.tar.gz

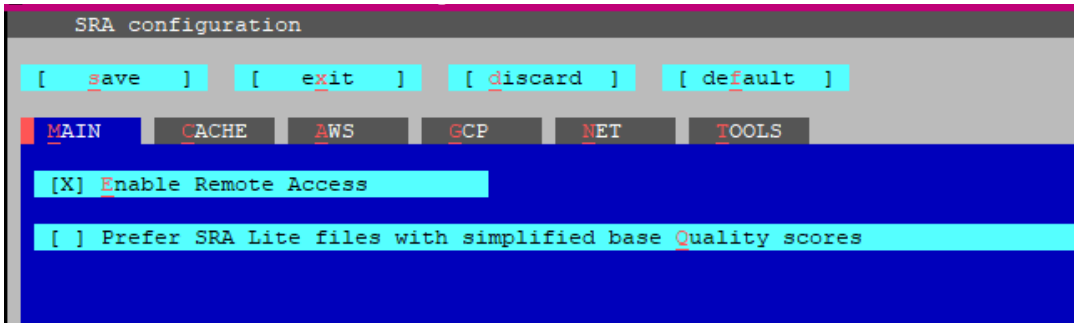
Instalación en MAC:

- pendiente

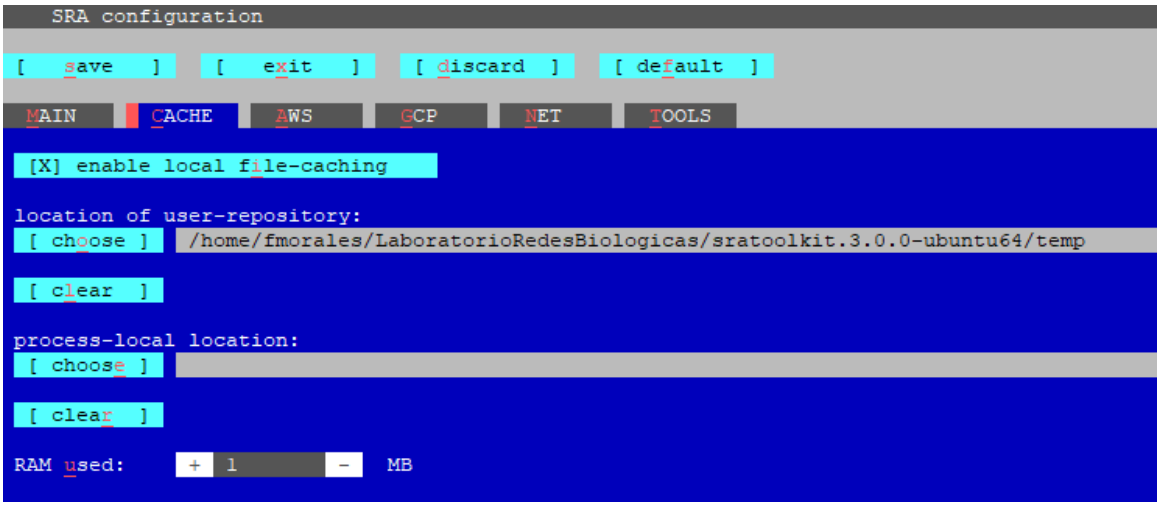
Configuración SRAtoolkit (Cualquier SO)

- Navegar hasta la ruta .../sratoolkit.3.0.0/bin, teniendo en cuenta la ruta utilizada en la instalación.
- Ejecutar el siguiente comando:

  ./vdb-config -i
- Se obtendrá una interfaz como la siguiente:



- Navegar hacia la pestaña CACHÉ



- Se debe configurar una ruta en el apartado "location of user repository". Esta será la ruta donde la función *prefetch* almacena los archivos que descarga de la base de datos NCBI. Sugerimos que la ruta especificada se encuentre dentro de la carpeta de sratoolkit y tenga el nombre de temp. Esto facilitará el manejo del código pero la elección final depende del usuario.
- No olvidar guardar los cambios antes de salir.

Nota: a veces la terminal de windows impide abrir la configuración, esto se soluciona ejecutando el símbolo del sistema (powershell o cualquiera que se esté utilizando) como administrador.

Para mayor información sobre la instalación y manejo de la herramienta SRAtoolkit, visitar la siguiente página:
https://github.com/ncbi/sra-tools/wiki

Flujo de Trabajo

Se tiene una notebook (TratamientoDeGenes.ipynb) con todo el código necesario para trabajar la cual se recomienda revisar de forma minuciosa antes de comenzar a hacer cualquier cosa. A grandes rasgos podemos dividir el flujo de trabajo en tres principales tareas:

- Recuperación de archivos de la base de datos NCBI (python)
- Conteo de Genes mediante R
- Normalización de datos (python)

Las primeras dos tareas son las más exigentes hablando de recursos computacionales (espacio y capacidad de procesamiento), por ello se recomienda realizarlas en el Servidor del Laboratorio. Así que se tienen dos códigos en python creados específicamente para ser ejecutados en el servidor (DescargaNCBI_Server.py y ConteoGenesR_Server.py). La última parte se recomienda hacerla en el Sistema Operativo de su preferencia y mediante el Notebook, esta sólo requiere de los archivos .csv generados en la etapa del Conteo de Genes. Sin embargo, de nueva cuenta, queda a elección del usuario el SO que utilizará para todo el flujo de trabajo.

Si se desea trabajar con el Servidor, las credenciales necesarias deberán ser solicitadas al Dr. Edgardo Galan. La conexión se realiza mediante protocolo SSH y esta puede

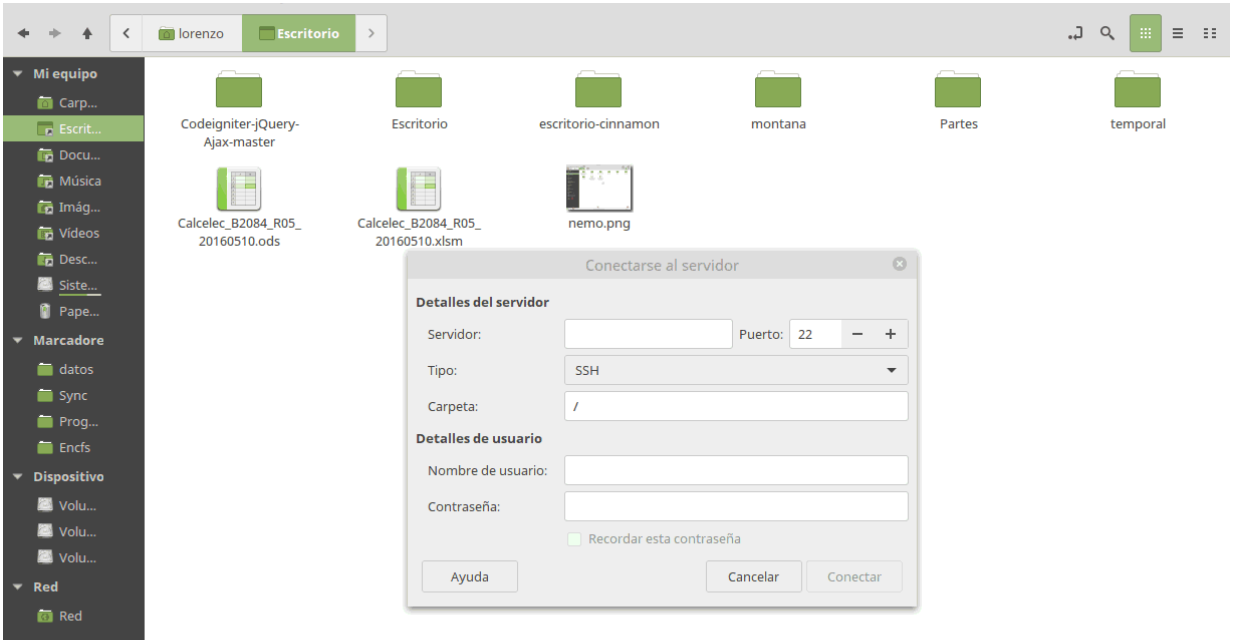
llevarse a cabo desde cualquier terminal mediante el comando:

ssh <usuario>@132.248.51.100 -p 2222

Recomendaciones para Linux:

Con la terminal del SO y su explorador de archivos es suficiente para trabajar en el servidor, facilitando enormemente la conexión y el intercambio de archivos.

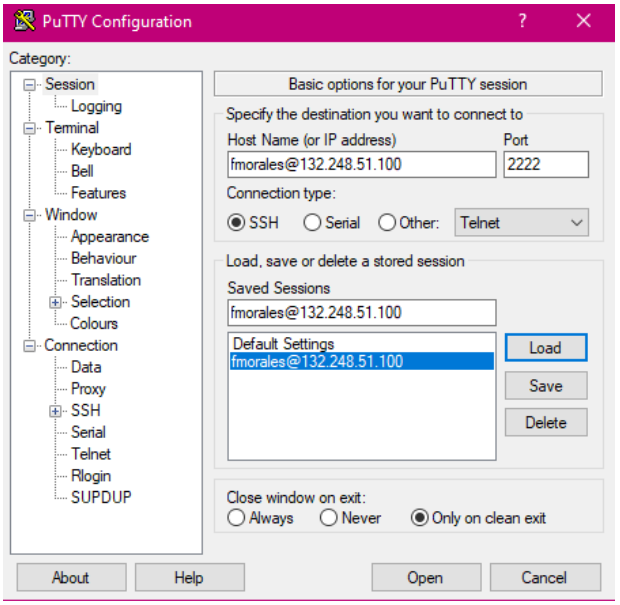
Ejemplo con Linux Mint: Sólo es necesario dar click en la opción de Red para obtener la ventana de conexión al servidor. Sólo restaría ingresar las credenciales solicitadas.



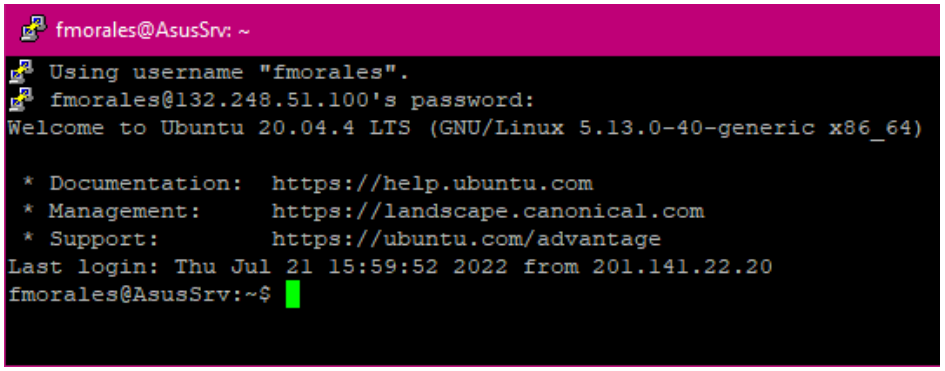
Recomedaciones para Windows:

Se recomienda el uso del software PuTTY para evitar los problemas de desconexión del servidor que suelen ocurrir si se utiliza el símbolo del sistema o powershell.

Download PuTTY - a free SSH and telnet client for Windows



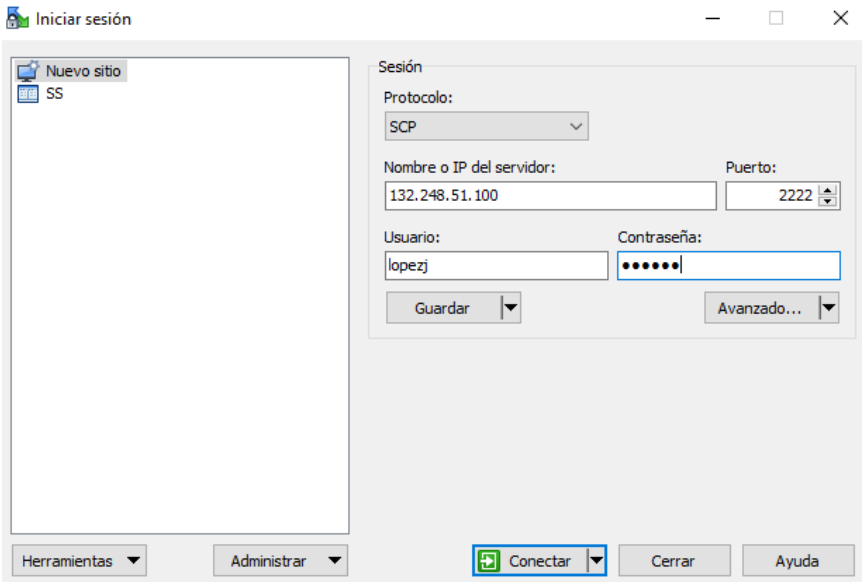
Ejemplo de Conexión mediante PuTTY



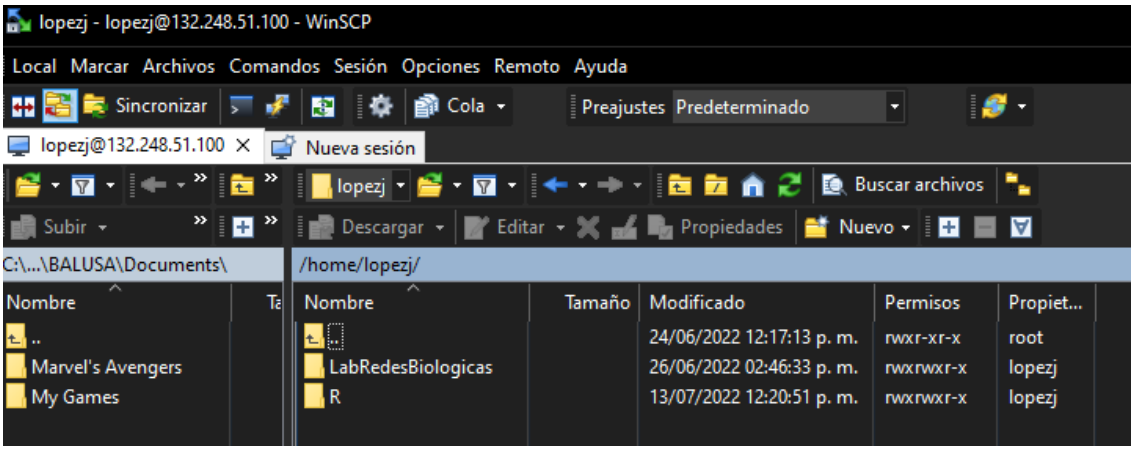
Interfaz de trabajo

Para el manejo de archivos se recomienda el uso del software WinSCP.

[WinSCP :: Official Site :: Download]



Ejemplo de Conexión mediante WinSCP



Interfaz de Trabajo

Estableciendo Espacio de Trabajo

Un aspecto de suma importancia a considerar antes de comenzar a trabajar, es el manejo de rutas. Para evitar errores se recomienda que todas las rutas utilizadas no tengan espacios o caracteres especiales.

En el código existe una sección donde se establece el entorno de trabajo. Es importante modificar ciertas líneas para trabajar de manera correcta y ordenada. El primer cambió (de ser requerido) es el número de plataforma que se trabajará. De igual forma, el indicar cuál será la ruta donde se guardarán todos los archivos y documentos será modificada en la línea work_dir. Otros cambios significativos son la identificación y colocación de las rutas para los

archivos del genoma de referencia, la carpeta de SRAtoolkit y el software R; cabiendo mencionar que estas últimas 2 rutas no son necesarias especificar si se trabaja con Linux.

Recuperación de archivos de la base de datos NCBI

En este apartado se realiza minería de datos para un organismo biológico asignado por el Dr. Edgardo. Dicho organismo tiene un número de plataforma asociado, que también es proporcionado por el Dr. A su vez, cada plataforma se divide en series de experimentos RNA-seq realizados por científicos e instituciones de cualquier parte del mundo, cada serie cuenta con un conjunto de *n* muestras y cada muestra puede presentar diferentes casos según sea la forma en que se llevó a cabo el experimento:

- Muestras Single (generan un sólo archivo)
- Muestras Paired (generan dos archivos)
- Varias "pasadas" (se generan varios archivos ya sean Single o Paired)

El código es capaz de manejar todas estas situaciones tanto para la descarga como para el conteo de genes, permitiendo una abstracción a nivel de muestras.
La recuperación de archivos se realiza mediante la herramienta SRAtoolkit haciendo uso de dos funciones: *prefetch* y *fasterq-dump*. La primera de ellas descarga los archivos necesarios de la base de datos y la segunda los transforma a formato fastq, formato especialmente diseñado para el almacenamiento y procesamiento de secuencias genéticas.

Si se trabaja con Windows es necesario ubicarse en el directorio *bin* para hacer uso de las funciones, aunque esta parte ya está especificada en el código por lo que no requiere un manejo más complejo. Por otro lado, si se trabaja con Linux es necesario agregar el directorio *bin* a las

variables de entorno en cada inicio de sesión, esto se hace mediante el siguiente comando:

export PATH=$PATH:$PWD/sratoolkit.3.0.0-<SO>/bin

Donde <SO> dependerá del SO utilizado y podemos saberlo simplemente viendo el nombre de la carpeta que se genera al instalar el SRAtoolkit. Por ejemplo para el caso del servidor que es Ubuntu, el comando sería el siguiente:

export PATH=$PATH:$PWD/sratoolkit.3.0.0-ubuntu64/bin

Para verificar que los binarios fueron encontrados, se ejecuta el siguiente comando:

which fastq-dump

Como salida se debe observar una ruta que corresponde a la función de *fastq-dump* perteneciente a las utilidades del SRAtoolkit.

```
fmorales@AsusSrv:~/LaboratorioRedesBiologicas$ which fastq-dump
/home/fmorales/LaboratorioRedesBiologicas/sratoolkit.3.0.0-ubuntu64/bin/fastq-dump
```

Ejemplo de salida del comando which

Recomendaciones:

- En el código se especifica la descarga para todas las series de experimentos asociadas a la plataforma del organismo, sin embargo, esto consumiría una ENORME cantidad de espacio, es por ello que se recomienda descargar y procesar serie por serie o conjuntos pequeños de series. La parte de código que se debe modificar está explícitamente señalada dentro del mismo mediante comentarios.
- El producto obtenido es un archivo en formato fastq el cual se comprime para ahorrar espacio. Se recomienda modificar el parámetro de tasa de compresión según sean las necesidades del usuario (ahorrar espacio o ahorrar tiempo), la especificación de este cambio

también se encuentra claramente indicada dentro del código.

Posibles Errores en el apartado de Descarga

Conteo de Genes Mediante R

Esta sección si bien está especificada en Python, hace uso de un Script en lenguaje R para llevar a cabo el conteo de Genes. A grandes rasgos, el programa toma los archivos *fastq* y les aplica un proceso de *trimming* mediante la herramienta SeqWin en el que se eliminan las secuencias de genes de baja calidad, la salida de este proceso es un archivo "*_trimmed.fastq".* Posteriormente se utiliza la herramienta RSubred para realizar el conteo de genes comparando los archivos *trimmed* y el genoma de referencia. El producto final es un archivo *.csv* con una matriz que representa el conteo de genes de cada una de las muestras de la serie.

Recomendaciones:

- Al igual que en la etapa anterior, se recomienda realizar el conteo de genes de forma individual para cada serie o en pequeños conjuntos.
- Si se cuenta con una buena capacidad de recursos computacionales a nivel de hilos de procesador y memoria ram, es posible acelerar el proceso de conteo de genes modificando los parámetros subReadThreads y shortreadRAM, el uso de estos parámetros se encuentra indicado dentro del código del Script de R (gene_count.R).

Posibles Errores en el Conteo de Genes

Recomendación antes de la etapa de Normalización:
Si al terminar la etapa de conteo de genes todo parece correcto y los archivos .csv se han generado correctamente, se recomienda eliminar todos los archivos asociados al conteo de genes: archivos trimmed, índices y bam. Esto con el objetivo de liberar espacio.

Si en el futuro se desea realizar de nueva cuenta algún conteo de genes u otro tipo de procesamiento se recomienda guardar los archivos fastq originales, en caso contrario se recomienda eliminarlos para liberar espacio.