Optimización de los parámetros de la herramienta de filtrado de alineamientos

Partimos del directorio “optimización” dentro del clúster. Cada usuario tendrá disponible este directorio en su directorio personal.

Tenemos dos scripts llamados: “run.sh” y “collect_data.sh”

También tenemos los subdirectorios de los transcriptomas de diferentes individuos, por ejemplo “TF36”.

Dentro de ese subdirectorio se encuentran otros 3:

- _Alineamientos_reformat: Contiene los ficheros mafft.fasta correctamente formateados en los cuales aún no se ha aplicado un filtrado.
- _SI: Contiene los ficheros maff.fasta que han pasado el filtro manual (criterio humano)
- _combinaciones: Contiene a su vez subdirectorios con lotes de combinaciones de parámetros para ejecutar el filtrado con todas ellas. Por ejemplo: “1-400”, “401-800”, etc.

Para iniciar el filtrado con todas las combinaciones de parámetros ejecutamos varias veces el primer script.

En cada ejecución le pasamos como primer argumento un subdirectorio con un lote de combinaciones (ej: TF36/TF36_combinaciones/1-400/) y como segundo argumento el directorio del individuo (ej: TF36/). Recuerda que entre el primer y segundo argumento hay un espacio en blanco.

Ejemplos de ejecución:

- sbatch run.sh TF36/TF36_combinaciones/1-400/ TF36/
- sbatch run.sh TF36/TF36_combinaciones/401-800/ TF36/
- sbatch run.sh TF36/TF36_combinaciones/801-1200/ TF36/

Podemos consultar las ejecuciones (o trabajos) del clúster con el comando “squeue”.

Para individuos con 2000 alineamientos tardará menos de 24 horas aproximadamente.

Una vez finalizado, dentro los directorios en los que se encuentran las combinaciones de parámetros se habrán creado dos directorios:

- Filtered_alignments: contiene los ficheros maff.fasta que han pasado el filtro automático.
- Coincidencias: contiene ficheros txt con información sobre la comparación entre el filtrado manual y automático.

Para recoger en una única tabla todos los datos ejecutamos el segundo script pasándole como argumento el directorio del individuo. Por ejemplo:

- sbatch collect_data.sh TF36/

La ejecución concluirá en cuestión de minutos. En el directorio del individuo obtendremos un fichero resultados.csv que recoge toda la información. Podremos descargarlo desde Bitvise y abrirlo con Excel.

Nota: para replicar todo este proceso con más individuos o muestras, es necesario añadir un nuevo directorio que contenga la misma estructura que TF36, es decir, dentro de él deben haber 3 directorios cuyos nombres terminen en "_Alineamientos_reformat", "_combinaciones" y "_SI".

El directorio "_combinaciones" se puede conseguir copiando "plantilla_combinaciones" (se encuentra dentro de "optimización") y cambiándole el nombre a uno más adecuado.

Si queremos unas combinaciones de parámetros diferentes, tendremos que crear una nueva plantilla con el script "combinaciones.py" que se encuentra dentro del directorio python_scripts

Dentro de combinaciones, los lotes pueden estar distribuidos de cualquier forma, por ejemplo, puede haber un único directorio (ej: "1-2000"), puede haber dos ("1-1000" y "1001-2000"), etc.

Por cada lote, se ejecutará una vez el script. Recuerda que en el clúster podemos lanzar varios trabajos para que se ejecuten simultáneamente.