Parallel BLAST

Important!

You have to create a public key for the remote host computer (a cluster, like Mazorka), for the par-blast to work we need to allow the host to access your computer without it asking for a password.

From your computer:

First create a directory named *parBlast* in your computer (use **mkdir**), move into it and create two more directories with the names *input_split_blast* and *blast_results_split*.

If there is already a *parBlast* directory make sure that, except for the *input_split_blast* and *blast_results_split* directories, there is nothing within it. Also, there are empty before running the parallel blast.

Concatenate all your *.faa* sequence files, this command should help:
`cat *.faa > Concatenados.faa`

Copy the *Concatenados.faa* file into the *par-blast* directory in your computer.

You have to use this file to create a database for blast:
`makeblastdb -in Concatenados.faa -dbtype prot -out Concatenados`
NOTE: the output file **must** be named *Concatenados* for the next scripts to work.

Now, we have to count the number of sequences within the file:
`grep '>' Concatenados.faa | wc`

Example: `grep '>' Concatenados.faa | wc`
`379570 379570 13469390`
This means that we have 379570 sequences in our file.

You have to divide this number between the number of splits you want, we usually make 16 splits. We need to find out the number of sequences per split (we will call this *yourNumber*). So, 379570 divided by 16 is 23724 approximately.

Run the script named *3.split_multifasta.pl*:
`./3.split_multifasta.pl -in=/mypath/parBlast/Concatenados.faa -output_dir=/mypath/parBlast/input_split_`
`--seqs_per_file=yourNumber`

We need to create a file that list the number of files that were generated by the split, go to the *input_split_blast* directory and use the next command:
`ls *.fsa | sort -g >fasta_files.txt`

Use **scp** to copy the *fasta_files.txt* file to your PARBLAST directory in your account on the remote host (the cluster).
`scp ~/mypath/parBlast/input_split_blast/fasta_files.txt user@remotecluster:~/mypath/PARBLAST/fasta_file`

From the Host computer (the cluster, like Mazorka).

Log into the host computer.
If a par-blast has been run before make sure to empty the *LOGS* directory in your home directory in your account at the host computer. If it doesn't exist, create it with **mkdir**.

Go to the PARBLAST directory and run the script named *4.blast.iMac-nodes.pl*:
`./4.blast.iMac-nodes.pl`

That's it

Now you only have to wait for your results to appear in the *blast_results_split* directory.