Tutorial 2 – A complete pipeline to process PHD (Phred quality) files in a cDNA sequencing project

Introduction

In this tutorial we will describe a pipeline that is identical to the one described in Tutorial 1, except that instead using trace files (chromatograms) as original input, we will employ PHD files. PHD are files produced by Phred program during the base calling and quality evaluation process. They are comprised by a header and three rows listing the DNA bases, the corresponding quality values and a scanning coordinate.

In order to upload PHD files instead of chromatograms, we will use the component `upload_phd_dir.pl`. However, why are we bothering with PHD files? The reason is that they are a good way to share sequence plus quality information across the Internet in a compact format. While chromatogram files generally occupy circa 220 kb (around 90 kb when compressed), uncompressed PHD files present only 9 kb. This means that a whole set of a genome project can be rapidly shared through the Internet.

EGene also presents another component (`upload_phd.pl`), which allows for uploading a single file containing concatenated PHDs. Thus, it is possible to concatenate numerous PHD files using UNIX `cat` command and then upload it directly to a pipeline created by EGene.

The following steps constitute this pipe:

1. Uploading PHD files;
2. Masking primer sequences;
3. Masking vector sequences;
4. Filtering low quality sequences;
5. Saving sequences invalidated by the quality filter;
6. Trimming the bases that present a low Phred quality value and those that are masked;
7. Filtering sequences considered too small;
8. Saving sequences invalidated by the size filter;
9. Filtering mitochondrial sequences;
10. Saving sequences invalidated by the contaminant filter;
11. Filtering ribosomal sequences;
12. Saving sequences invalidated by the contaminant filter;
13. Saving sequences not previously invalidated by any filter;
14. Generating a report of all filtering steps;
15. Creating an XML snapshot recording all the processing steps that were performed;
16. Assembling the valid sequences using CAP3;
17. Generating an HTML page with graphical reports;
18. Generating a complete graphical report.

We have previously constructed a pipeline for this tutorial using CoEd, EGene's graphical configuration editor. The EGene's configuration file (`phd_complete.gen`) and its counterpart text file (`phd_complete.cnf`) can be found at the `config_files` directory. In order to run the pipeline, go to the `/examples/phd_complete_pipe` directory. This directory contains the subdirectory `phd_dir`, which presents a set of PHD files, and the file `primer_table.txt`, composed by a list of the primers used in the sequencing.

To run the pipe, you should type the following command:

```
bigou.pl -c ../config_files/trace_file_complete.cnf
```

If everything goes well, you should now find the following additional files in this directory:

```
filtered_by_quality.fasta
filtered_by_size.fasta
filtered_by_mitochondria.fasta
filtered_by_ribosome.fasta
filtering_report.html
redundancy_report.html
report_graphic_simple.html
final_snapshot.xml
good_sequences.fasta
```

and the following additional directories:

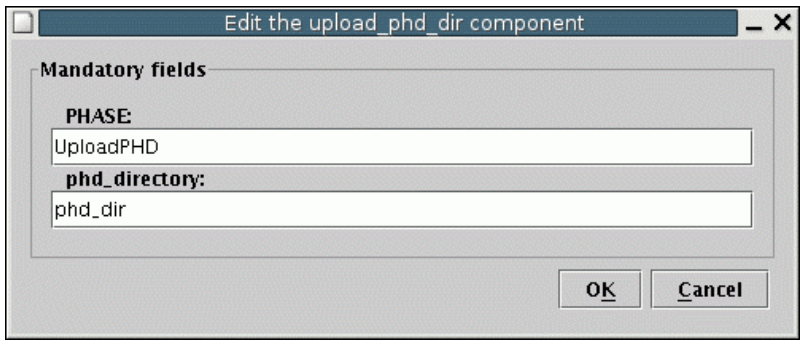
```
assembly_dir/
complete_report/
images_dir/
```

Only the first component of this pipeline will be discussed in this tutorial, as all the other steps are identical to those described in Tutorial 1. Please refer to it in case of doubts. The results obtained in this example should be also identical to those of Tutorial 1, except that the directory `/assembly_dir/chromat_dir/` will not contain symbolic links to the trace files, since they are absent in this case.

1. Uploading PHD files

Configuration parameters in the .cnf file:

```
#==============================================================
PHASE=Upload PHD
program = upload_phd_dir.pl
#--------------------------------------------------------------
directory = phd_dir
#==============================================================
```



This step uses the component `upload_phd_dir.pl` to upload a set of PHD files located at the directory specified by the user. The only argument to this component is the directory that contains the PHD files (in our case `phd_dir`). It is assumed that `bigou.pl` is run while the shell is in the directory that contains `phd_dir`. Alternatively, the user can specify a complete path for the directory (e.g. `/home/test/phd_dir`).