**Shaghayegh Hosseini email:Shaghayegh.hosseini@chalmers.se**

Abstract

In the area of bioinformatics, developing computational tools and pipelines are a crucial part to draw biological conclusion out of “raw data” coming form Sequencing machines. The purpose of this project is to develop a python based pipeline that automates data processing of 80 different *H.Pylori*.

Raw data is in so called FastQ format with average size of 1 Gigabyte. FastQ is a text-based [format](http://en.wikipedia.org/wiki/File_format) for storing both a biological sequence and its corresponding quality scores, which are encoded with a single [ASCII](http://en.wikipedia.org/wiki/ASCII) character for brevity.

Data processing starts with pre processing which usually includes quality control and data trimming. The next phase is to feed the quality controlled reads (short biological sequences) into assembly software, which matches the reads to generate the whole biological sequence, in this project, genome sequence of *H.Pylori*. The final step is to look for genes in genome and map them with genes in public databases with the aim of annotation (assigning biological functions to each gene).

Problem and solution

There is already a Perl based pipeline, which handles this process.

The problem is that we have 80 different datasets that for each of them some sub processes for example the process of quality control and trimming can be done in parallel and independently. The existing pipeline does this process in a sequential fashion. The solution here is to use **Parallel programming**. Then the process will take much less time.

Time plan

As illustrated in figure 1, first three days will cover the implementation part, which includes converting code from Perl to python, using Bio python modules and parallel programing. Day fourth is to test the pipeline with some pilot datasets and day fifth is for finalizing the report.

Figure 1 Project time plan