**Final Project Proposal**

**Analysis of a sequence using Bowtie 2 for aligning sequence reads to reference genome.**

**Objective:**

Bowtie2 is a fast and efficient tool for aligning sequence reads to long reference reads. It indexes the genome based on the Burrows-Wheeler Transform algorithm. Since it is the first step in almost every NGS analysis pipeline it would be useful to have a web-based analysis application for this tool.

**Proposal:**

The web-based analysis application will require components like user interface that will allow the user to upload or select FASTQ files. The frontend HTML page will have JavaScript and JQuery client-side interaction. The UI will have the functionality for the user to choose between a single or a paired library for the analysis. It will then show one or two file browse boxes depending on the earlier selection. Depending on the library selection the user will be required to upload one (for single) or two (for paired) read files. The UI will also show a dropdown box for selecting the reference genome for the analysis. For the reference genome, the drop-down will show a static list of genomes. The user will select the reference genome to be used for the analysis. Later on, this selected genome reference will be used to query a database to retrieve the reference FASTA file. The UI will also show fields related to bowtie options and specifications. The bowtie usage section will provide user with options under bowtie2 aligner. The options will include default, end-to-end read alignment and paired alignment. Under the default mode the script will include only the read file(s) and indexed reference. For end-to-end and paired alignment, the user will have the option to select from some of the preset commands that are provided by the tool. Some of the presets included in the end-to-end mode will be --fast, --sensitive and for local mode --fast-local, --sensitive-local.

The UI will use a form element to upload the file as well as other inputs to the server. Form validation using JavaScript and JQuery will be carried out for, a) the input file that needs to have fastq extension. The user will be shown an error message in case files with an incorrect extension are selected. b) all mandatory fields will be checked for non-empty values.

Once the user submits the form, the request will be sent to webserver where the CGI and HTML template will handle the backend. The CGI script will make an ftp connection to the NCBI database to download the user selected reference genome file. The CGI script will also be responsible to run the bowtie -build function on the selected reference genome files creating the index files. This will be run as a default setting and the user will not be provided with options. The script will also execute the bowtie2 script with user selected bowtie align options. This will result in an output file in the default SAM format. The resulting alignment rate and other specifics like alignment type (paired/unpaired), alignment mode (default, end-to-end, paired), assembly source (human, mouse) will be stored in a table schema which will later be read and displayed in a visual form. The styling of the output results will involve CSS. The schema for storing this information will have an assembly table with alignment Id (AlnID) as the primary key. A new entry will be given a unique AlnID by parsing the table and incrementing the previous AlnID number. The table will also have an assembly source, as same assembly source can have different alignment summaries based on the read input but each analysis will use only one reference sequence. The alignment summary table will include the summaries for each unique alignment Id.

