CGAL - 0.8.0 BETA Manual

1 Introduction

CGAL is a tool for comparing genome assemblies. It computes the likelihood of a set of paired end reads given an assembly which can be used as a metric for evaluating the assembly.

2 Installation

CGAL can be run on any environment that supports POSIX threads. To install it execute the following commands:

```
make
```

3 Running

To run CGAL the reads need to be mapped to the assembly using an external mapper. At present we support Bowtie 2 and BFAST. The output of the mapper need to be in SAM format. Once mapping is done, running CGAL consists of three steps:

- The first step converts the output generated by Bowtie 2 or BFAST to an internal format.

  If you are using Bowtie 2, please use '-a –no-mixed' options to map and then run

  ```
  ./bowtie2convert out.sam maxFragmentLength
  ```

  If BFAST is used run

  ```
  ./bfastconvert out.sam maxFragmentLength
  ```

  where `out.sam` is the output file of the mapper in SAM format and `maxFragmentLength` is the maximum length of fragment. Mappings with longer fragment length than `maxFragmentLength` will be ignored. If this value is set too high intermediate files will be quite large and runtime will be greatly increased. The default value is `5000`.

- The tools for mapping are usually not able to align all the reads. To align the reads not mapped by the mapper, we have adapted the striped implementation of Smith-Waterman algorithm by Farrar. However, this step is time consuming. So, we align only a random subset of reads.

  ```
  ./align contigFile toAlign numThreads
  ```

  where `contigFile` is the assembly file in FASTA format, `toAlign` is the number of reads to be aligned and `numThreads` is the number of POSIX threads.

- cgal This step computes the likelihood value.

  ```
  ./cgal contigFile
  ```

  where `contigFile` is the assembly file in FASTA format. The ouput is written to file `out.txt`.

3.1 Notes

- Each step generates intermediate files used by later steps. These files must not be deleted until final step is completed.
- CGAL does not filter out reads other than the ones with more than 80% of the read covered with N's.

4 Output

The ouput file contains following values separated by tabs:

- Number of contigs
- Total likelihood value
- Likelihood value of reads mapped by the mapping tool
- Likelihood value corresponding to reads not mapped
- Total number of paired-end reads
- Number of reads not mapped by the mapper