Consensus Module Manual

Karin A. Remington

1. Overview

The Consensus Module (CNS) generates a consensus sequence for each contig in the layout by first multi-aligning the read fragments and surrogates contained within the contig and then assigning a base to each position based on some weighted scoring of the appropriate column in the multi-alignment. The initial release performs naïve multi-alignment and consensus scoring, while subsequent implementations will provide an refinement process to improve the multi-alignment with scoring based on the quality values of the reads.

2. Memory Usage and Performance

Memory requirements for the Consensus module scale linearly with contig size. A typical instantiation of the 1/1000th Human simulation (a004.sim) results in Nuggetizer contigs of maximum length 50Kbp. Consensus processes its input on a contig by contig basis, and so is limited by this maximum contig size. At 10x coverage, the memory required is on the order of 10 Mb. In the case of the full Human genome assembly, consider 500Mbp to be the length of the longest possible contig (based on an conservative estimate of the length of the longest Human chromosome). At 10x coverage, we expect 5 Gbp of read data to process. To include both base and quality-value data then requires at least 10Gb of memory.

Preliminary timing results show that processing for the 1/1000th Human simulation requires approximately 1 cpu minute, and 1/100th Human requires just over 10 cpu minutes. This demonstrates the expected linear scaling of execution time based on problem size.

3. Command Line Interface

Command line interface for prototype:

```
consensus [-q] [-r] FragStorePath CGWStream
```

`-q` requests that quality values be used to break ties in calling bases (initial behavior is tobreak ties randomly)

`-r` Levels 2 and 3: requests that alignments be refined via pass through realignment step.

The `FragStorePath` directs the module to the location of the persistant fragment store generated by previous stages of the Assembly system. The `CGWStream` consists of messages of type `SurrogateMesg` and `IntContigMesg` from the Assembly module (as described in the CNS I/O specfication document), with the ordering requirement that any `SurrogateMesg` referenced in a given `IntContigMesg` appears in the stream prior to its reference.

AUTHORS

Karin A. Remington:

`Created February 12, 1999`

`Modified March 11, 1999`

```
   Modified May 10, 1999 (added -r option, -p becomes default)
```