Parallelizing BWA Using Work Queue and Hadoop MapReduce

Christopher Ray
CSE 40822
Background: What is Ancient DNA?

- DNA recovered from biological specimens that has not been preserved specifically for later analysis
- DNA isolated from ancient specimens

Image from: http://news.psu.edu/story/141655/2010/01/19/research/mammoth-achievement-researchers-forefront-molecular-biology
Main Idea

- Plant debris that falls to lake eventually settles on lakebed, creating layers
- Ancient DNA samples found in deeper layers of lakebed should be older than those found above
- If you determine what species are present at each layer, you can trace how the plants in a surrounding area changed over time

Image from: http://bactifeed.com/biological_soil_conditioner_FAQ.html
Problems

- Burrows-Wheeler Alignment Tool (BWA) is slow for large queries
  - ~23 minutes to align 8 GB fastq query
  - >1 hour to align 31 GB
- Using BWA’s multithreading feature can speed up runtime, but can lead to varying results compared to running sequentially

http://bio-bwa.sourceforge.net/
Solutions: Work Queue and Hadoop MapReduce

Images from: [http://bactifeed.com/biological_soil_conditioner_FAQ.html](http://bactifeed.com/biological_soil_conditioner_FAQ.html) and [http://ccl.cse.nd.edu/software/workqueue/WorkQueueLogoSmall.png](http://ccl.cse.nd.edu/software/workqueue/WorkQueueLogoSmall.png)
Work Queue Model

WQ_bwa (Work Queue master)

Reference.fas
Query.fq

sam_cleanup

0.fq, 1.fq, 2.fq ...

Worker output

Condor Pool
Worker 1
Worker 2
Worker 3
Worker 4
Worker 5
...

output.sam
Using Hadoop MapReduce

- Uploaded query fastq file to HDFS
- Ran the following command:

```
```
## Results

<table>
<thead>
<tr>
<th></th>
<th>Runtime (seconds)</th>
<th>Speedup</th>
<th>Efficiency (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>1282</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>WQ (50 workers)</td>
<td>463</td>
<td>2.769</td>
<td>5.538</td>
</tr>
<tr>
<td>WQ (100 workers)</td>
<td>406</td>
<td>3.158</td>
<td>3.158</td>
</tr>
<tr>
<td>WQ (150 workers)</td>
<td>403</td>
<td>3.181</td>
<td>2.121</td>
</tr>
<tr>
<td>WQ (200 workers)</td>
<td>404</td>
<td>3.173</td>
<td>1.587</td>
</tr>
<tr>
<td>MapReduce</td>
<td>332</td>
<td>3.861</td>
<td>6.657</td>
</tr>
</tbody>
</table>
Challenges

- Implementing Work Queue model that continuously submitted/waited for tasks in the queue rather than submitting all tasks at once
- Configuring optimal Work Queue task size
- Configuring BWA to be usable with Hadoop MapReduce
Future Work

- Continue configuring ideal task size
- Possibly implement method for choosing different BWA algorithms within Work Queue Model and Hadoop MapReduce