Hardware Acceleration of the Short Read Genome Reassembly Problem

Carl Ebeling, Scott Hauck, Larry Ruzzo, Maria Kim, Corey Olson, Cooper Clauson, Boris Kogon, and Mark Seligman

Acknowledgements: Pico Computing, Washington Technology Center (WTC)
Introduction to DNA Sequencing

- Nucleotide bases: A, T, G, C
- Encodes information for various applications

Human Genome Project

- Officially took 13 years
  - 1990-2003
- 3 billion base pairs
- $3 billion project

Source: http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml
Sequencing Goals

• The 1000 Genomes Project
  • Studying diseases
  • Personalized medicine

• $1000 per genome

• Next Generation Sequencing technology

Source: http://bioinformatics.genomics.org.cn/image/1000genomes.jpg
Next Generation Sequencing—4 steps

1. Replicate

2. Slice randomly

3. Sequence

Source: http://www.illumina.com/images/technology/sequencing_system_workflow_e_lg.jpg
Next Generation Sequencing—4 steps

4. Alignment

ACTCAC GACGTTAA
ACTCAGGTGACGTTAA
ACTCAGGTGACGTTAA

Reference sequence

Reads

ACTCAGGTGACGTTAA

Reads- to sequence

Reference- already sequenced
The Short Read Genome Reassembly Problem

• Polymorphisms
• Repeats
• Insertions/deletions
• Sequence errors
The Short Read Genome Reassembly Problem

• Expensive computation
  • 200 million 30-200 base pair short reads
  • 3 billion base pair reference
Our Solution

1. **Hardware platform**

2. Optimized algorithm
Hardware Background Information

- FPGA—Field Programmable Gate Array
- Reconfigurable hardware
• Parallel computation blocks
  ❖ Logic
  ❖ DSP
Pico Computing
Our Solution

1. Hardware platform

2. Optimized algorithm
The Smith Waterman Algorithm

- Inexact Matching
- Computationally expensive
- Key: Smith Waterman on selective areas

<table>
<thead>
<tr>
<th></th>
<th>G</th>
<th>C</th>
<th>C</th>
<th>C</th>
<th>T</th>
<th>A</th>
<th>G</th>
<th>C</th>
<th>G</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Reference</strong></td>
<td><strong>Read</strong></td>
<td><strong>Reference</strong></td>
<td><strong>Read</strong></td>
<td><strong>Reference</strong></td>
<td><strong>Read</strong></td>
<td><strong>Reference</strong></td>
<td><strong>Read</strong></td>
<td><strong>Reference</strong></td>
<td><strong>Read</strong></td>
</tr>
<tr>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>C</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>A</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>T</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>G</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

Source: http://www.ibm.com/developerworks/java/library/j-seqalign/LCSTab5.gif
Algorithm

1. Find Candidate Alignment Locations (CALs)
2. Run Smith-Waterman on the Read and its CALs
## Determining the CALs

### Reference Index Table (RIT)

<table>
<thead>
<tr>
<th>Reference</th>
<th>Index 1</th>
<th>Index 2</th>
<th>Index 3</th>
<th>Index 4</th>
<th>Index 5</th>
<th>Index 6</th>
<th>Index 7</th>
<th>Index 8</th>
<th>Index 9</th>
<th>Index 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAA</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAAAAC</td>
<td>38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATGCCGC</td>
<td>5</td>
<td>23</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATGCCGG</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATGCTAA</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATGCTAC</td>
<td>8</td>
<td>19</td>
<td>63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATGCTAG</td>
<td>3</td>
<td>26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTTTTTTTT</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

---

The reference sequence is highlighted in purple, and the CALs (CATGCTAC) are marked in blue.
Using the Reference Index Table

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Index 1</th>
<th>Index 2</th>
<th>Index 3</th>
<th>Index 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAAAA</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAAAAAAC</td>
<td>38</td>
<td>88</td>
<td></td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>CATGCCGC</td>
<td>5</td>
<td>23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATGCCGG</td>
<td>0</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CATGCTAA</td>
<td>12</td>
<td>37</td>
<td>786</td>
<td></td>
</tr>
<tr>
<td>CATGCTAC</td>
<td>8</td>
<td>19</td>
<td>63</td>
<td>155</td>
</tr>
<tr>
<td>CATGCTAG</td>
<td>3</td>
<td>26</td>
<td>362</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>TTTTTTTTT</td>
<td>45</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Segment from a read

**CALs:** 8, 19, 63, 155
Block Diagram

- CAL Finder
- Smith Waterman Unit
- FPGA
- RAM (Table + Reference)

Connections:
- Read
- {Read, [CALs]}
- {Read, Reference_Location}
Block Diagram—Parallelize

- CAL Finder
- Smith Waterman Controller
- Smith Waterman Array
- RAM (Table + Reference)
- FPGA

Read

{Read, [CALs]}

{Read, Reference_Location}
Block Diagram – Optimize

CAL Finder → CAL Filter → Smith Waterman Controller → Smith Waterman Array

FPGA

{Read, [CALs]}

{Read, Reference_Location}

RAM (Table + Reference)
Performance

- BFAST: 6-7 hours on a 16 core machine
- Our estimated performance: 35 seconds on 9 FPGAs
Summary

- DNA sequencing assists in many applications
- Next generation sequencing
- The short read genome reassembly problem
- FPGAs and our algorithm accelerate the computation