Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and study.

2013-11-18 Lab meeting

Yifan Yang
• Background
• Scientific question
• Algorithm the paper introduced
• Discussion
PacBio® Data Characteristics

Depending on the insert size of the library, the PacBio® RS can be optimized to generate longer sequences or shorter but higher quality sequences.

**Continuous Long Reads (CLR)**. A large insert library (e.g., 6-10kb) results in long CLR reads up to 10 kb:

![Continuous Long Reads Diagram]

**Circular Consensus Sequence (CCS)**. A short insert library (e.g., 500-1000 bp) favors multiple passes around each circular SMRTbell™ construct. The sequence generated by multiple observations of a single DNA molecule can be summarized as a higher quality (>99% accuracy) consensus sequence.

![Circular Consensus Sequence Diagram]
Background: Advantage v.s disadvantage of PacBio SMS

Disadvantage:
PacBio Long reads: ~11% error (15%)
Error: uniformly distributed insertions and deletions, very few substitution

Advantage:
mean = 2246bp
max = 23,000bp
Single molecule sequencing
Available for RNA-seq-long reads RNA-seq

Two main problems of short reads RNA-seq:
Ambiguous reads mapping
Assembly for transcripts

Will PacBio SMS be a future?
Question: How to map the reads to genome with high error rate?
3 steps Algorithm

Find anchor

rough alignment

refinement alignment
Two critical questions:

1. How to define the length of anchor K?
2. Whether anchor number N will influence mapping?
How to define the length of anchor K?
How many bases have to be at least sequenced to get a K error-free length?
They expect to find at least 10 anchors for one read

X: number of anchors
Y: probability of sequencing at least X anchors.

L = 1000
Similar regions need more anchors.
As the read length increases, the mapping quality increases!
Simulated datasets

**Figure 8.**
"Real" datasets btw different methods

Table 2
A comparison of the BLASR, BWA-SW, and BLAT methods on _E. coli_ reads

<table>
<thead>
<tr>
<th>Method</th>
<th>Number of aligned reads</th>
<th>Number of aligned bases</th>
<th>Run time</th>
</tr>
</thead>
<tbody>
<tr>
<td>BLASR-SA</td>
<td>94057</td>
<td>230.8 M</td>
<td>20m 54s</td>
</tr>
<tr>
<td>BLASR-BWT</td>
<td>94527</td>
<td>230.1 M</td>
<td>33m 57s</td>
</tr>
<tr>
<td>BWA-SW</td>
<td>97729</td>
<td>132.4 M</td>
<td>434m 5s</td>
</tr>
<tr>
<td>BLAT</td>
<td>99530</td>
<td>181.7 M</td>
<td>4724m 40s</td>
</tr>
</tbody>
</table>

Table 3
A comparison of the BLASR, and BWA-SW methods on simulated reads

<table>
<thead>
<tr>
<th>Method</th>
<th>Correctly mapped reads</th>
<th>Correctly mapped bases</th>
<th>Incorrectly mapped reads</th>
<th>Incorrectly mapped bases</th>
<th>Skipped reads</th>
<th>Skipped bases</th>
<th>Runtime</th>
<th>Memory footprint</th>
</tr>
</thead>
</table>
| _E. coli_
| BLASR-SA | 108739                 | 255.5M                 | 223                      | 0.39M                    | 3766          |              | 49m 16s | 202 MB           |
| BLASR-BWT| 108735                 | 255.3M                 | 259                      | 0.45M                    | 3604          |              | 59m 39s | 46 MB            |
| BWA-SW   | 111192                 | 251.9M                 | 1825                     | 0.91M                    | 3003          |              | 223m 57s| 190 MB           |

| H. sapiens
| BLASR-SA | 41726                  | 102.3M                 | 1074                     | 1.89M                    | 413           |              | 92m 26s | 14.7 GB          |
| BLASR-BWT| 41582                  | 101.7M                 | 1159                     | 1.75M                    | 472           |              | 53m 26s | 8.1 GB           |
| BWA-SW   | 40981                  | 96.3M                  | 292                      | 1.16M                    | 1554          |              | 105m 24s| 4.2 GB           |
Discussion

• How to evaluate method?
• Long reads rna-seq is promising?
Thanks!