Computational Challenges in NGS Data Analysis and the Way Forward

2nd National Workshop on Marker-Assisted Selection for Crop Improvement, ICRISAT, Patancheru, 28 October, 2010

Asoke K Talukder, Ph.D
Geschickten Solutions, Bangalore, India
Structure of the Talk

• Introduction
• Challenges
• Complexity of the Problem
• Complexity in Extracting Results
• Deep Computing (HPC and Cloud Computing) - the Way Forward
• Q&A
Credits & Acknowledgements

- This presentation draws upon many excellent images and literatures prepared by others and freely available on the Web. This artwork is being used here solely for nonprofit academic purposes.

- To all the anonymous authors and artists!
Complexity

Can Computing & Algorithms help prove the hypothesis?

http://anyg.cn/GUEST/
Omic Data
Scientific Opportunity
Comprehensive, field-scale understanding of the responses of soil microbial communities and their processes to long-term (10 yr) elevated CO$_2$, and comparison of responses across terrestrial ecosystems.
Qualitative to Quantitative Biology

- Better Predictability
- Higher Accuracy
- Less time to market
Proprietary Tools

• Offered by Equipment vendor
• Available to Equipment owners (not to users!)
• General Purpose, Tested, and Maintained
• User Friendly with Good GUI
• Works on Microsoft Windows
• Limited Scope
• If it does not work – BAD LUCK!!!!!
• NO CUSTOMIZATION
Open Domain Tools

• Developed by Ph.D/Post Doc students
• One Problem – many Solutions
  – everyone claims to be best
  – which one is the best?
• Available to Public Free of cost
• Generally developed in UNIX and C/C++
  (algorithmic portion) and Java (visuals part)
• Tested for specific Ph.D problem, may not be rigorously tested, and NOT Maintained
• Tools are data depend – does not work all types, shapes, or size of data
Complex Problems in Biology

1. Genome Assembly
2. Metegenomic Assembly
3. Gene Hunting within a Genome
4. Gene Prediction/Annotation from cDNA, rRNA, EST data
5. Understanding Static & Dynamic Properties of Molecules
6. Molecular Sequence Alignment & Variation Analysis
7. Search various Databases & Knowledgebases
8. Integration & Systems Biology
9. Evolutionary tree estimation
10. Discovering novel patterns and target molecules
11. Discovering Combinatorial properties of biological networks
12. Simulating Systems Biology space
13. ::
14. ::

• Each of these is addressed through algorithms to solve a computational problem, typically based upon optimizing some mathematical criterion
Quality of Results

• As our knowledge of Biology is limited
  – We do not exactly know how much to expect
  – The data generated from experiment are not 100% complete & trustworthy
  – NP-hard problems can lead only to approximate heuristic solution

• It is like using a map (without the orientation) for a journey where we do not know “where we want to go?”, “where we are?” or “where to start?”
Most Biology Solutions are NP-Complete & Algorithm Design

- If the data volume increases by x, complexity of solution is much higher than x (non deterministic polynomial)
- Getting exact solutions may not be possible for some problems on some inputs, without spending a great deal of time
- You may not know when you have an optimal solution, if you use a heuristic
- Almost impossible to arrive at exact solution; however, if the solution is obtained, it can be proved it is the right solution
- Sometimes exact solutions may not be necessary, and approximate solutions may suffice. But, how good an approximation does the solution need?
Sequence Assembly

• Almost any *in-silico* problem has Assembly as its first step, be it
  – Genome Assembly
  – Transcriptome Analysis
  – Metagenomics Study
  – EST Assembly
  – SNP Analysis
  – ChIP Sequence Analysis
Sequence Assembly
The Sequencing & Assembly Process

1. Biological Sample (DNA)
2. Break the Large Molecule into Tiny Fragments
3. Sequence (Read) Each Fragments
4. Align Overlapping/Contiguous Sequences (Assembly Process)
5. Nature
6. Wet Lab
7. Dry Lab
8. Multiple genomes
9. Random genomes fragmentation
10. Target Microbial Genomes
11. Genomes assembly using overlaps
Challenges

- **Contamination error**
- **Chimeric fragment error**
- **Base-call errors** occur at rates 1 to 5 errors every 100 nucleotide
- **Computational errors**

- **How to validate** – what you got is right?
Steps in Assembly

• Understand the data
  – Data inventory
  – Single End, Paired End, Mate Paired etc
  – Sequence structure (Read size, Format)
  – Quality of the data

• Clean up the data
  – Remove (Filter/Trim) vector/adaptor contaminated data
  – Remove data of bad quality
  – Remove data that might cause chimeric error

• Assembly & Downstream Data Analysis
Statistical Analysis of Data

Figure 1: Total number of reads per lane across all flow cells

Figure 2: Total number of tiles per lane across all flow cell

Figure 3: Number of FASTQ records per lane per flow cell

Figure 4: Percentage of best matched reads per lane per flow cell:
Computational Errors
(Repeated Regions)

Ref: Joao Setubal, Joao Meidanis, Introduction to Computational Molecular Biology
Computational Errors
(Repeated Regions)

Ref: Joao Setubal, Joao Meidanis, Introduction to Computational Molecular Biology
Computational Errors (Identity Vs Similarity)

- Nature is not an Innovator
- Nature is a TWEAKER!
- So we do not look for identity but similarity
- Which similar part?
  - How much Similar?
  - Same motif seen repeating in many places, which is correct?
Computational Errors
(Chimeric Error)

• Unrelated reads are connected for short reads
  – Poly A or Poly T will result in Chimeric error during RNA-Seq or Metagenomic Assembly
  – Repeat regions will result in Chimermic error in DNA Assembly
  – Error in Read will cause chimeric errors during Assembly
  – Error in Read/Sequences will change the course of analysis for a sRNA
Denovo Genome Assembly

Velvet is widely used for Denovo assembly

Selecting K-mer length is one of the most important & complex step in Denovo assembly process.

Less k-mer length creates higher data utilization and more contigs with large Graphs – Needs huge memory (RAM)

Larger k-mer uses lesser data, creates lesser contigs, but needs lesser RAM

Velvet has used 143 GB of RAM to run with the k-mer length of 41 on Sequences on 42 Million reads of read length 100 for a domestic animal
## Denovo Assembly: Bacteria

<table>
<thead>
<tr>
<th>K-Mer Length</th>
<th>Coverage Cutoff</th>
<th>No. of contigs</th>
<th>Min. Contig Length</th>
<th>Max. Conting Length</th>
<th>No. of Nucleotides</th>
<th>N50</th>
<th>N90</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>5</td>
<td>41201</td>
<td>37</td>
<td>3528</td>
<td>6968557</td>
<td>251</td>
<td>76</td>
</tr>
<tr>
<td>19</td>
<td>15</td>
<td>31993</td>
<td>37</td>
<td>3528</td>
<td>6849824</td>
<td>328</td>
<td>95</td>
</tr>
<tr>
<td>25</td>
<td>5</td>
<td>7336</td>
<td>49</td>
<td>22860</td>
<td>6559364</td>
<td>2257</td>
<td>586</td>
</tr>
<tr>
<td>25</td>
<td>15</td>
<td>2606</td>
<td>49</td>
<td>55461</td>
<td>6445341</td>
<td>7121</td>
<td>1795</td>
</tr>
<tr>
<td>31</td>
<td>5</td>
<td>4240</td>
<td>61</td>
<td>30906</td>
<td>6532619</td>
<td>4372</td>
<td>1209</td>
</tr>
<tr>
<td>31</td>
<td>15</td>
<td>1420</td>
<td>61</td>
<td>80206</td>
<td>6440483</td>
<td>15672</td>
<td>3796</td>
</tr>
<tr>
<td>37</td>
<td>5</td>
<td>2474</td>
<td>73</td>
<td>50637</td>
<td>6503211</td>
<td>8763</td>
<td>2365</td>
</tr>
<tr>
<td>37</td>
<td>15</td>
<td>990</td>
<td>73</td>
<td>84326</td>
<td>6440421</td>
<td>22469</td>
<td>5777</td>
</tr>
<tr>
<td>43</td>
<td>5</td>
<td>1305</td>
<td>85</td>
<td>77722</td>
<td>6468483</td>
<td>18483</td>
<td>4817</td>
</tr>
<tr>
<td>43</td>
<td>15</td>
<td>738</td>
<td>85</td>
<td>112109</td>
<td>6437558</td>
<td>33114</td>
<td>7733</td>
</tr>
<tr>
<td>49</td>
<td>5</td>
<td>677</td>
<td>97</td>
<td>94660</td>
<td>6440466</td>
<td>29516</td>
<td>8134</td>
</tr>
<tr>
<td>49</td>
<td>15</td>
<td>609</td>
<td>97</td>
<td>118347</td>
<td>6433474</td>
<td>32766</td>
<td>8755</td>
</tr>
<tr>
<td>55</td>
<td>5</td>
<td>624</td>
<td>109</td>
<td>94597</td>
<td>6442218</td>
<td>31231</td>
<td>8634</td>
</tr>
<tr>
<td>55</td>
<td>15</td>
<td>559</td>
<td>109</td>
<td>119105</td>
<td>6435216</td>
<td>33891</td>
<td>9964</td>
</tr>
<tr>
<td>57</td>
<td>15</td>
<td>578</td>
<td>113</td>
<td>120216</td>
<td>6433931</td>
<td>31608</td>
<td>8454</td>
</tr>
<tr>
<td>59</td>
<td>15</td>
<td>599</td>
<td>117</td>
<td>118680</td>
<td>6432860</td>
<td>27122</td>
<td>7496</td>
</tr>
<tr>
<td>61</td>
<td>15</td>
<td>623</td>
<td>121</td>
<td>98495</td>
<td>6426299</td>
<td>24008</td>
<td>6952</td>
</tr>
<tr>
<td>67</td>
<td>15</td>
<td>1143</td>
<td>133</td>
<td>1503</td>
<td>333921</td>
<td>301</td>
<td>190</td>
</tr>
<tr>
<td>73</td>
<td>15</td>
<td>267</td>
<td>145</td>
<td>720</td>
<td>67227</td>
<td>258</td>
<td>168</td>
</tr>
<tr>
<td>79</td>
<td>15</td>
<td>44</td>
<td>157</td>
<td>532</td>
<td>9831</td>
<td>198</td>
<td>167</td>
</tr>
<tr>
<td>85</td>
<td>15</td>
<td>5</td>
<td>179</td>
<td>287</td>
<td>1206</td>
<td>275</td>
<td>179</td>
</tr>
<tr>
<td>91</td>
<td>15</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Denovo Assembly: Bacteria
1 sample of Genomic data (Marine Animal)

No. Of reads in End1 = 77,195,083
No. Of reads in End2 = 77,195,083
Total no of reads = 154,390,166 (154.3 Million)
Total no of bases = 9,263,409,960 (9.26 Billion)

Time taken to run TopHat on 24 Core Intel Xeon Processor with 144 GB RAM is 33h 25m 2 sec

Total no of Conserved Genes produced by running TopHat considering data as paired end = 1074

Total no of Conserved Genes produces by running TopHat considering data as Single end = 1367
Entropy in SNP Analysis

Alignment statistics by different software

Biology data will give different reference assembly on different run

This is worse with low coverage

<table>
<thead>
<tr>
<th>Software</th>
<th>Total Reads</th>
<th>Aligned Reads</th>
<th>% of Aligned Reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAQ</td>
<td>4,575,106</td>
<td>944,514</td>
<td>20.64%</td>
</tr>
<tr>
<td>Bowtie</td>
<td>4,575,106</td>
<td>918,244</td>
<td>20.07%</td>
</tr>
<tr>
<td>BWA</td>
<td>4,575,106</td>
<td>833,466</td>
<td>18.2%</td>
</tr>
<tr>
<td>Mosaik</td>
<td>4,575,106</td>
<td>837,352</td>
<td>18.3%</td>
</tr>
</tbody>
</table>
Coverage (SNP)
Unused Data & Entropy

Unused Data in velvet for Domestic animal 77.29%  
(31829820 Unused reads out of 41,179,668) – increases  
with higher k-mer value

Unused Data in MAQ for Domestic animal with a similar  
species as reference Genome 74.76% (31,456,340  
Unmaped reads out of 42,071,130)

Unused Data in TopHat for a marine animal in reference with  
Genome_super_Contigs 98.5% (70,934,677 Unused  
reads out of 71,954,940)

2 Million sequences of small RNA gave potentially 20,000  
(~1 %) sequences suitable for miRNA stem-loop search
Metagenomics Challenges

- No separate genome or transcriptome sequencing
- Enormous datasets with high gene density
  - Large compute resources required
  - 2 orders of magnitude jump
- Fragmentary data
  - Inadequate bioinformatics tools for assembly, annotation, analysis, visualization
- Metadata standards non-existent
  - Metadata absent from databases
  - Lack of standards impedes collection of datasets
- Diversity of User Sophistication and Needs

Source: JGI, USA
Human Architecture! Growth Performance

Source: Rajkumar Buyya
Cloud Computing

- When you need milk, you don’t need to buy a Cow
- As a biologist, when you need to do some experiment, should you be bothered about bioinformatics or a supercomputer and the maintenance of the terabytes of storage?
Cloud Computing Defined

- Cloud computing is an emerging computing paradigm where data and applications reside in the cyberspace, it allows users to access their data and information through any web-connected device be it fixed or mobile.

- Source: John B. Horrigan, Use of Cloud Computing Applications & Services, Data memo, PEW Internet & American Life project, September 2008
## Computational Power Improvement

### No. of Processors

<table>
<thead>
<tr>
<th></th>
<th>C.P.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

### Computational Power Improvement

- **Uniprocessor Supercomputers** (tall, taller)
- **Multiprocessor** (fat, fatter)

Source: Rajkumar Buyya
Biology Problems, HPC & the Cloud

• As Computers are becoming fatter (multiple cores) and clusters are becoming cheaper, it is slowly becoming possible to attempt solving NP-complete problems in Biology
• All computing algorithms to solve biology problems must be parallel & distributed
• Cloud computing will play a significant role in this attempt
Potential Solution – Way Forward

• Combine and leverage the advancements of Computing Technologies like HPC
• Efficient and Optimized Algorithms
• Interdisciplinary team of Computer Scientists, Biologists, Mathematicians, Statisticians and HPC experts

GenomicsCloud
We all (?) use The Cloud

Benefits of the Cloud

• Helping Green computing by lending out idle resources through Cycle Scavenging

• Unlimited Resource
  – Unlimited Computing power
  – Unlimited storage (Filestore & online memory)
  – Scale UP or Scale Down On-demand

• Users can use resources without owning anything
  – converting Capex to Opex

• Research Labs & Small & Medium Biotech companies can use Supercomputers/HPC without owning them

• Pay as you go on Job/task at hand
Regulatory Requirements

• Genomic Data related to Human cannot go offshore
• Some metegenomic data are too defense critical data; therefore, cannot go offshore
• Some genomic data are too confidential to be processed in a lab offshore
• Data need to be Processed & Analyzed onshore that needs tools, infrastructure, and expertise
GenomicsCloud Architecture

- **SaaS**: Cloud-MAQ, Velvet/P, NGSdacs
- **PaaS**: Phoenix (Middleware)
- **IaaS**: HPC Cluster, Storage, Servers
Reference Assembly on the Cloud

Cloud-MAQ: The Cloud-enabled Scalable Whole Genome Reference Assembly Application

Asoke K Talukder*¹, Santhosh Gandham*², Prahalad H. A*³ and Nitai Pada Bhattacharyya*⁴

<table>
<thead>
<tr>
<th></th>
<th>Paired End, 50 base length</th>
<th>hg19.fa</th>
<th>28h 38m 11sec</th>
<th>Singlenode / 4core Intel Xeon, 8GB RAM</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAQ</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cloud-MAQ</td>
<td></td>
<td>hg19.fa</td>
<td>9h 36m 59sec</td>
<td>Singlenode / 4core Intel Xeon, 8GB RAM</td>
</tr>
</tbody>
</table>
Welcome...Johnson

"DEEP COMPUTING FOR DEEP SEQUENCING"

Cloud-MAQ

Cloud-MAQ is a parallel, scalable, Cloud-Computing ready Mapping and Assembly with Quality which increases the performance of MAQ reference assembly multi-fold. The input for this analysis is Read data and Reference.

- Ref Assembly
  - Cloud-MAQ
  - MPI-MAQ
+ Denovo Assembly
+ Statistical Analysis
+ mRNA Analysis
+ micro Array
+ BLAST
Thank you

Asoke K Talukder, Ph.D
Adjunct Faculty, ABV Indian Institute of Information Technology, Gwalior
Adjunct Professor, National Institute of Technology, Warangal
Adjunct Faculty, National Institute of Technology, Warangal
Ex DaimlerChrysler Chair Professor, IIIT Bangalore
Email: “asoke” dot “talukder” (at) “geschickten” dot “com”