Novel HPC technologies for Rapid Analysis in Bioinformatics

Presenter: Paul Walsh, nSilico Life Science Ltd, Ireland

Tristan Cabel, Gabriel Hautreux, Eric Boyer: CINES, France

Simon Wong: ICHEC, Ireland

Nicolas Mignerey: GENCI, France
Novel HPC technologies for Rapid Analysis in Bioinformatics

- Synthetic Biotech Revolution!
- Revolutionizing and becoming an integral part of every major industry, including medicine, energy, agriculture and manufacturing.
The computational challenge is to find often subtle patterns in Peta Bytes of nucleotide and protein data.

NCBI has 30 Peta Bytes Archived
Novel HPC technologies for Rapid Analysis in Bioinformatics

Central Dogma of Biology

DNA → Transcription → RNA → Translation → PROTEIN

DNA sequence:

```
GTGCATCTGACTCCTGAGGAGAAG
CACGTAGACTGAGGACTCCTCTTCC
```

RNA sequence:

```
GUGCAUCUGACUCUCUGAGGAGAAG
```

Protein sequence:

```
VHLTPEEK
```
Novel HPC technologies for Rapid Analysis in Bioinformatics

- In **bioinformatics**, a sequence alignment is a way of arranging the sequences of **DNA**, **RNA**, or **protein** to identify regions of similarity that may be a consequence of functional, **structural**, or **evolutionary** relationships between the sequences.

  ![Sequence Alignment Example](http://en.wikipedia.org/wiki/Sequence_alignment)

  * symbol shows where proteins are identical

http://en.wikipedia.org/wiki/Sequence_alignment
Novel HPC technologies for Rapid Analysis in Bioinformatics

- Sequence Alignment
- Why? Assess similarity of sequences and learn about their relationship, evolutionary, functional, structural, etc..

\[
\text{Sequences} \quad \text{Alignment}
\]

\[
\begin{align*}
&\text{ACCCGA} \\
&\text{ACTA} \\
&\text{TCCTA} \\
\Rightarrow \text{align} \\
&\text{AC--TA} \\
&\text{TCC--TA}
\end{align*}
\]

Complexity can be \(O(n^2)\)

Novel HPC technologies for Rapid Analysis in Bioinformatics

- HPC more urgent to service sequencing advances.
- Advances in sequencing technology for genomics, transcriptomic, proteomics...
Novel HPC technologies for Rapid Analysis in Bioinformatics

• We are partnering with Dell to provide affordable and secure private cloud

• Direct access to data via Illumina Base Space
Novel HPC technologies for Rapid Analysis in Bioinformatics

- Sequencing is getting cheaper but ........
- Bioinformatics has a high cost

<table>
<thead>
<tr>
<th>Sample collection &amp; experimental design</th>
<th>from blood samples (easy to collect) to brain tissue (hard to collect)</th>
<th>~$100 onwards from a few hours to several days</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequencing</td>
<td>Library preparation + running the sequencer (whole dual flow cell)</td>
<td>~$6500 = ~$500 + ~$6000</td>
</tr>
<tr>
<td></td>
<td>~380M reads/lane; 1 individual; ~1140M total reads (~3 lanes for a 30x coverage); ~250Gb (intermediate files)</td>
<td>~11-12 day</td>
</tr>
<tr>
<td>Data reduction &amp; management</td>
<td>Alignment (transfer* and storing raw data + mapping)</td>
<td>~$40 = ~$33 + ~$7 300Gb (BAM file)</td>
</tr>
<tr>
<td></td>
<td>(data transfer and storage for 10 days)*; **</td>
<td>~1/2 day *** (including transferring 250Gb FASTQ ~7.5 hrs)</td>
</tr>
<tr>
<td></td>
<td>SNP calling (compute + transfer out)</td>
<td>&lt;$5 = ~$4 + ~$0.60 &lt; 1Gb</td>
</tr>
<tr>
<td></td>
<td>Indel calling (compute + transfer out)</td>
<td>&lt;$35 = ~$32 + ~$0.60 &lt; 1Gb</td>
</tr>
<tr>
<td></td>
<td>SV calling (compute + transfer out)</td>
<td>&lt;$35 = ~$32 + ~$0.60 &lt; 1Gb</td>
</tr>
<tr>
<td>Downstream analyses</td>
<td>&gt;$100K</td>
<td>~310Gb months</td>
</tr>
</tbody>
</table>

Over $100,000 !!!
Novel HPC technologies for Rapid Analysis in Bioinformatics

- HPC
  - “bioinformatics tools for processing and analyzing data from NGS are relatively new, and in many cases not well adapted for HPC”
Novel HPC technologies for Rapid Analysis in Bioinformatics

• Our software aims to provide Speed, Simplicity, Security, Scalability
  – Design with biologists for biologists
  – Specialized low cost energy efficient hardware
  – Full security and traceability
  – Use High Performance Computing in a Cloud Computing Framework (public/private) with Map Reduce

Need to explore novel HPC strategies
Novel HPC technologies for Rapid Analysis in Bioinformatics

Simplicity Microbio

Just upload sequence and click go for rich report

Name of analysis: **Analysis of Staphylococcus aureus**

Abstract:

Infections (e.g. boils), respiratory disease (e.g. sinusitis), and food poisoning. Disease-associated strains often promote infections by producing potent protein toxins, and expressing cell-surface proteins that bind and inactivate antibodies. The emergence of antibiotic-resistant forms of pathogenic S. aureus has highlighted the pressing need for effective new treatments.

Upload sequence data and we will automatically configure the best analysis

- [Dropbox](#)
- [Other File Sharing System](#)

Bacteria Genome assembly and annotation for Illumina data

Trimming Option (Cutadapt)

- Read Type: ○ Paired end ○ Mate pair ○ Unpaired
- Trim adapter from reads? ○ yes ○ no

SynBio 2014
Sample file:
- Reads Library type: Paired end
- Filename ERR064898_2.fastq.gz
- File type Conventional base calls
- Encoding Sanger / Illumina 1.9
- Total Sequences 1,913,937
- Sequence length 75
- %GC 33

Analysis shows high similarity in function to sub-species Z712
Novel HPC technologies for Rapid Analysis in Bioinformatics

- With map reduce and multi-threaded programming we time to generate analysis from 140 hours down to 9 hours.
- Embarrassingly parallel approach.
- Need to extract further efficiencies

SHAPE PROVIDED
MUCH NEEDED
EXPERTISE

Need to explore novel HPC strategies for alignment
Novel HPC technologies for Rapid Analysis in Bioinformatics

- Public and Private Cloud Implementations are available with Parallel Distributed Processing
- Numerous algorithms were benchmarked for accuracy and speed.
- Our first challenge is to parallelise the Smith-Waterman local alignment algorithm.
- We needed expert assistance to help us achieve this.
- PRACE SHAPE support has made this possible
Novel HPC technologies for Rapid Analysis in Bioinformatics

- Smith-Waterman alignment algorithm: used by many bioinformatics packages (e.g. short read mapping).

- PRACE experts identified the initial codebase and built parallelised versions of the Smith-Watermann algorithm, with many-core technology in mind.

- Successfully applied for compute resource allocation on MareNostrum at BSC (hybrid nodes).

- Version developed for current x86 architectures.

- Another version also developed for future Intel architectures (Haswell and Knight Landing).
Novel HPC technologies for Rapid Analysis in Bioinformatics

- How to improve the Smith Watermann algorithm?

Now:

- USE INTEL XEON PHI VECTORIZATION
- PARALLELIZE ON THE COUPLED READ/DATABASE
Novel HPC technologies for Rapid Analysis in Bioinformatics

• How to improve the Smith Waterman algorithm?

In a near future:

NEW AVX2 = INCREASED COMPUTE CAPABILITIES

INTENSIVE PARALLELISM FOR BETTER PERFORMANCES

SynBio 2014
Novel HPC technologies for Rapid Analysis in Bioinformatics

- Theoretical results on 16-bits elements

<table>
<thead>
<tr>
<th>Instruction set</th>
<th>Number of instruction</th>
<th>Number of elements per vector</th>
<th>Theoretical speed-up</th>
</tr>
</thead>
<tbody>
<tr>
<td>SSE2 (128-bits)</td>
<td>53</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>AVX2 (256-bits)</td>
<td>66</td>
<td>16</td>
<td>1.60</td>
</tr>
<tr>
<td>Phi Corner (512-bits)</td>
<td>53</td>
<td>16*</td>
<td>2</td>
</tr>
</tbody>
</table>

* Because on Phi corner we have to use 32-bits scores

- Preliminary results on AVX2

  **22% speed-up** of the computation code using AVX2 (on a Haswell Pre-Production processor Xeon® E5-2697 v3)

- Expected results on **Intel Knight Landing®**

  High parallelism and AVX2 => about **50% of speed up**
Novel HPC technologies for Rapid Analysis in Bioinformatics

- OpenMP Scaling

On Intel(R) Xeon(R) CPU E5-2670 0 (8 cores) with
- Test case 1: database made of small elements
- Test case 2: database made of big elements
Novel HPC technologies for Rapid Analysis in Bioinformatics

• Outcome: valuable lessons and documentation for many-core parallelisation/porting work. Codebase ready for future processors. Follow-up work under discussion.

• International collaboration with complementary skills:
  – NSilico: bioinformatics and business insight, example datasets
  – EPCC: preparation of SHAPE application
  – ICHEC (Ireland): project coordinator and bioinformatics expertise
  – GENCI & CINES (France): many-core development expertise
  – BSC (Spain): Intel Xeon Phi hybrid nodes on MareNostrum
Novel HPC technologies for Rapid Analysis in Bioinformatics
Final thought

- We are keen to push the boundaries of what is possible in life science using HPC.
- We warmly welcome discussions on potential collaborations.

Contact: paul.walsh@nsilico.com