Cloud-Based High-Throughput Comparative Analysis of Plant Species Genomes

Samart Wanchana [1], Sarah Covshoff [2], Rowan Sage [3], Gane Wong [4], Julian M. Hibberd [2] and Richard M. Bruskiewich [1,5,6]

[1] International Rice Research Institute, Los Baños, Laguna, Philippines
[2] Dept. of Plant Sciences, Cambridge University, Cambridge, Cambs. UK
[3] Dept. of Ecology and Evolutionary Biology, University of Toronto, ON, Canada
[4] Dept. of Biological Sciences, University of Alberta, AB, Canada
[5] Dept. of Molecular Biology & Biochemistry, Simon Fraser University, BC, Canada
[6] Dephinai, Port Moody, BC, Canada

Correspondance: richard.bruskiewich@delphinai.com

The development of relatively low cost high throughput short read sequencing technology gives expanding opportunities for large scale inter-specific comparisons. At the same time, the sheer size of such data sets overwhelms even the best equipped of locally hosted computing facilities.

Fortunately, the emergence of large scale commercial and academic cloud computing facilties can help small research teams cope with the deluge of data. Even so, the processing of large, unstructured but relatively homogeneous data sets requires novel strategies and algorithms for parallel computation across many available virtual cloud instances.

Here we present a simple case study in parallel cloud based bioinformatics analysis, as applied to comparative genomics data consisting of 96 large transcriptome data sets from the 1000 plant genome project (http://www.onekp.com/). We also briefly mention the potential of cloud computing facilities for other areas of crop research informatics.

Our presentation focuses less on the specific results of this project , which are to be published elsewhere, than on the specific analytical strategies deployed on leased Amazon Web Services infrastructure. In particular, we show how algorithms like MapReduce, as implemented in the Apache Hadoop software system, can be used to solve specific tasks in this kind of research project. The general utility of leased cloud infrastructure for small research teams is also overviewed.

Materials & Methods

Source of the Project Data

RNA samples were prepared from juvenile and mature leaf tissues from a total of diverse C3 and C4 plant species grown in the Dept. of Ecology and Evolutionary Biology, University of Toronto, ON, Canada. These RNA samples were subsequently shipped to the Beijing Genome Institute, Shenzhen, PRC for sequencing using Solexa or Illuminex NGS, generating fastq files which were assembled by the BGI bioinformatics team into candidate transcriptome contigs using the SOAP tool suite. The resulting datasets were transferred to TACC, where BLASTX was used against several available databases of annotated genes which included plant genome refseq collections from NCBI (courtesy of JLM). The resulting blastx output files served as the initial raw inputs to the orthologous gene family analysis.

Orthologous Gene Family Identification

Parsing out of Best Hits

A Perl script using the BioPerl library was coded to parse out all 96 blastx result files. This processing was run on a high CPU(8 core) AWS EC2 instance, to generate key-value pairs representing the best hitsfor each assemble gene contig. Only ?efseqanchored hits were parsed out. When available, the best Arabidopsis refseq served as the key value. Alternately, the best Orzya sativa ssp. *Japonica* refseq hit was used. If neither was available, then any available hit was randomly used as the key. The values of the key pairs from a given blastx result consisted of the blast matches with the highest E value. A filtering threshold of 10e-4 was used to exclude lower quality hits.

Aggregation of Best Hits into Orthologous Gene Sets

A Hadoop MapReduce script was developed to aggregate all key-value best hits into orthogous gene clusters anchored on the key.

Hadoop run on 14-02-2011... with a 20 node cluster. Set MapReduce tasks to 40 since each instance of the cluster has 2 CPU's. Set jvm reuse to 砥 nlimitedin the blind faith that the overhead of computation will be large.

hadoop-0.20.2]$ sudo $HADOOP_HOME/bin/hadoop jar ~/ComparaGene-0.1.jar OrthologSets **-D mapred.reduce.tasks=40 -D mapred.job.reuse.jvm.num.tasks=-1** /user/root/besthits.out output

Curious observations:

- Map was fast, but reduce phase slows down remarkably at 67% completion. Why?
- Combiner input record number is 1.64 times larger than Map output record number

Acknowledgments

The raw BLASTX analysis results were kindly provided by Jim Leebens-Mack (jleebensmack@plantbio.uga.edu) as part of the transcriptome sequence post-processing analysis by NESCent and the iPlant Collaborative (at TACC) for the 1000 plant transcriptome project.