Contribution ID: **71** Type: **Presentation**

A bioinformatics user point of view of cloud computing

Tuesday, 18 September 2012 14:40 (20 minutes)

Description of the work

I will list some recent activities that have been computationally demanding and their computing environments. In the presentation I will include some wider examples beyond our own work.

-The SUPERFAMILY database pipeline (supfam.org) runs in the AWS EC2 cloud environment. We use this to process the results of genome and other sequencing. The flexibility allows us to use a large number of instances to return results quickly whilst being idle most of the time. Our demands fluctuate wildly, favouring such a flexible funding model. Furthermore the rate at which sequence data is being produced is rising faster than Moore's law, so we see cloud computing as the most scalable solution for the future.

-To give context to genome sequencing we must build phylogenetic trees of nearly two thousand species. The search space of a binary tree of this size is effectively infinite, but by using heuristic maximum likelihood algorithms we are able to calculate a biologically reasonable topology. Investigations of these trees were carried out on the Compute Canada national resource, accessing the order of 10^7 CPU hours.

-Next generation sequencing technology presents a significant computational challenge in assembling the overlapping raw fragments of data that are produced, into contiguous sequence. We carried out the assembly of about 70 full human transcriptomes, each requiring a minimum of 100GB of RAM taking many days. We also attempted an assembly of the combined set using 1TB of RAM. This was conducted on the Institute of Cancer Research HPC facility in London.

-Without going into details, constructing protein coding genes out of assembled DNA or RNA sequence is another challenge facing bioinformatics in the face of the deluge of next generation sequence data. We have written prototype software in Hadoop, developed on a local cluster, for effectively carrying out protein analysis directly on assembled DNA/RNA effectively bypassing the need for gene prediction.

Link for further information

http://www.cs.bris.ac.uk/~gough

Wider impact of this work

The motivation for presenting at this forum is twofold: first to give attendees (by describing use cases) an insight into the computational requirements of bioinformatics, the ways in which it is currently being deployed in different computing environments and give some orientation of the field of bioinformatics and the way it moves; secondly to bring back from the meeting information to disseminate to the community of bioinformatics developers and users. As a member of three influential consortia in bioinformatics, various biotechnology/medical national funding panels and international conference committees, I will contribute to bringing these applied areas forward in their use of cloud computing.

The wider impact is thus in influencing both communities and accessing, via this forum a channel of two-way communication. Promotion of cloud computing in the life sciences and raising awareness of bioinformatics to the developers of software and infrastructure will aid coordinated growth.

Printable Summary

Our research group has been making use of HPC and distributed computing for bioinformatics since 1999. We committed the majority of our research to cloud computing at the start of 2010, and see this as our long-term future.

This presentation will describe the various bioinformatics use cases in our research group and some from the wider field of bioinformatics in general. This includes e.g.: DNA and protein sequence analysis, phylogenetics, next generation sequence assembly, personal medicine, etc. The presentation will also describe some of the computational resources we use for bioinformatics including e.g.: AWS EC2 cloud, Compute Canada national distributed resource, local Hadoop, and national HPC.

This presentation is aimed at those interested in hearing stories of use cases for cloud computing in bioinformatics, and potential directions in which the field could be moving in the future with respect to its computational requirements.

Primary author: GOUGH, Julian (University of Bristol)

Co-authors: Mr SMITHERS, Ben (University of Bristol); Dr VAVOULIS, Dimitris (University of Bristol); Mr OATES, Matt (University of Bristol)

Presenter: GOUGH, Julian (University of Bristol)

Session Classification: Providing cloud services

Track Classification: Virtualised Resources: challenges and opportunities (Michel Drescher: track leader)