CEES bioinformatics – yearly report for 2014

Introduction

Bioinformatics – "an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data" (Wikipedia) continued to grow at CEES in 2014. An increasing number of researchers are generating large 'digital' datasets that need to be analyzed using sophisticated bioinformatics tools. We use bioinformatics in a wide sense here, so that the following types of researchers are examples of bioinformaticians at CEES:

- all users of 'R'
- those doing statistical modeling
- those analyzing next generation sequencing data, be it genomics, transcriptomics, or other such data
- those working with time series data
- those extensively using the unix command-line, programming in perl, python, C, etc.

People

Based on the number of researchers subscribed to the different mailing lists (see below) we estimate more than 50 CEES members substantially rely on bioinformatic methods during their research. A new trend for 2014 is the increasing number of master students making use of the resources for their projects.

Infrastructure

Strategic considerations

Many researchers are still able to perform their analysis on their desktop/laptop machine. However, increasingly, we need to perform tasks that are either CPU or memory intensive.

At CEES, we use a combination of self-owned servers, and CPU hours we applied for nationally on the UiO supercomputer 'Abel'. This maximizes flexibility for CEES researchers in choosing the right resource for their project:

- memory-intensive applications can be run on our own servers
- CPU-intensive applications can be submitted to Abel and therefore do not take up valuable time on the servers

The servers CEES owns (see below) are physically co-located with, and attached to, the Abel system. This means users can seamlessly access the same programs and disks on the self-owned servers, as well as on Abel.

For storage ('project disk space) we rent space from USIT at UiO (attached to Abel), rather than buy and administer our own. The benefit of this strategy is that we do not have to spend valuable research time on basic system

administrative tasks and software installation. The HPC (high-performance computing) group of USIT is very proficient in installing programs with difficult dependencies or requirements. Finally, backup of valuable research data is arranged for as well.

Hardware

Starting with the project to sequence and assemble the genome of Atlantic cod in 2009, CEES has invested in its own hardware for computation. These servers are hosted and maintained by the HPC group of USIT. In 2014, the following computational infrastructure was available to the CEES:

- two high-memory servers with 24 CPUs and 128 GB of RAM, and around 1 TB disk space each ('cod1' and 'cod2', bought in 2009)
- two high-memory servers with 64 CPUs and 512GB of RAM, and around 24 TB disk space each ('cod3' and 'cod4', bought in 2011)

The following resources were rented or allocated to CEES:

- on the University computer cluster (called 'Abel') we have an allocation for CPU-intensive computations. This allocation was successfully renewed and now amounts to 1.3 million CPU hours per half year. The allocation is usually insufficient but we can ask for an extension. We estimate we used around 4.4 million CPU hours in 2014.
- we rent 40 TB of project disk space (administrated by USIT, includes backup). The costs of this storage (1600 NOK/TB/year) are shared between users, proportionally to the used space
- we have, shared with the Norwegian Sequencing Centre, 30 TB disk space for long-term archival of data at norstore, the national Norwegian infrastructure for the management, curation and long-term archiving of digital scientific data

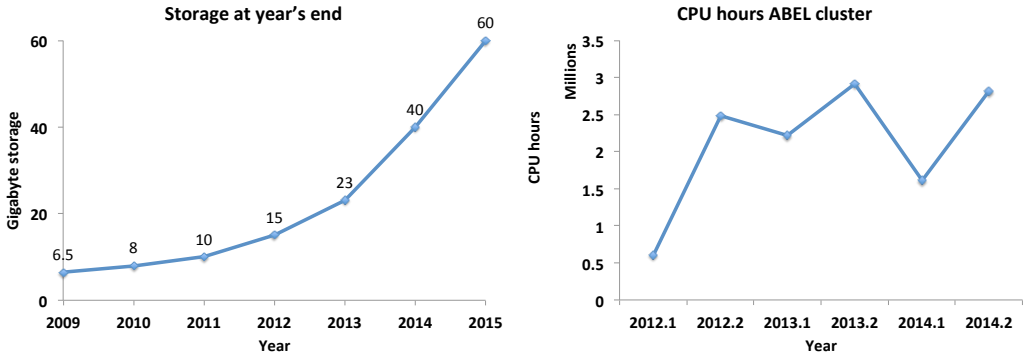

Figure 1: (left) rented storage at year's end (in gigabyte) and (right) CPU hours used per half-year period (in million CPU hours) by CEES HPC resource users

Investments in 2014

With support from the faculty of Mathematics and Natural Sciences, and in collaboration with the Norwegian Sequencing Centre and the Aqua Genome project, we were able to order hardware to completely renew as well as extend the CEES computer resources. From early 2015, we will have access to three high-memory servers with 1.5 TB RAM and 64 CPUs. These servers come with considerable local disk (64 TB per server). In addition, we will have access to

224 regular 'grid' CPUs (providing us with the equivalent of close to 2 million CPU hours).

Software

There are a few project specific applications available through the 'cod' servers:

- the program 'Stacks' for analysis of RAD-tag sequence data
- mysql servers for databases
- the web portal ('SMRTportal') for analysis of data from the Pacific Biosciences instrument

Administration

The day-to-day administration of the servers and disk space is the responsibility of USIT. However, there is still a considerable overhead for CEES staff:

- correspondence with USIT on required software, interruptions of the servers, feedback to CEES users
- keeping an eye on the disk space allocation (when the maximum is reached, new files cannot be written anymore without warning)
- communications with the users necessary for a smooth running of the shared resource (e.g., asking users to clean up disk space)
- administration of the user base, mailing lists etc.
- instructing new users, who often are new to the field

Projects

Types of projects requiring large computational resources and large amounts of disk space are:

- projects to generate de novo genome assemblies
- resequencing projects (SNP finding and genotyping) that require many CPU hours for mapping sequencing data to a reference
- metagenomics and environmental sequencing projects
- the RAD-seq platform at CEES (SNP detection and genotyping by sequencing): computations through the Stacks software on the cod1 server
- transcriptomics analysis pipelines
- sequencing of mitochondrial genomes
- projects using Illumina and Pacific Biosciences data to sequence and assembly bacterial genomes

Mailing lists

There are several mailing lists:

- cees-bioinf@bio.uio.no (59 subscribers, up from 45 in 2013), for general information exchange
- cees-hpc@bio.uio.no (38 subscribers, up from 25) for specific information regarding shared Abel resources
  cod.nodes@.bio.uio.no (40 subscribers, up from 26) for reserving the shared cod servers

Meetings

The bioinformaticians at CEES meet regularly through a new group started in 2013, called 'the Genome Analysis Club' (TGAC). At the meetings, papers and programs focusing on analytical methods for whole genome data (including whole genome sequencing, SNPs, RADtags, etc) are discussed.

Wiki

We use a UiO wiki with articles dealing with the practicalities of using the resources at CEES, tips and trick, etc. The wiki is open to the world (https://wiki.uio.no/mn/bio/cees-bioinf).

Courses

No local course activity took place in 2014. An effort was started to increase the offering of Software Carpentry workshops at UiO. From the website: 'Software Carpentry helps researchers be more productive by teaching them basic computing skills. We run workshops at dozens of sites around the world, and also provide open access material online for self-paced instruction. The benefits are more reliable results and higher productivity: a day a week is common, and a ten-fold improvement isn't rare.' With support from the faculty of Mathematics and Natural Sciences and the Science Library, the first workshop is planned to take place in February 2015 and will be open to all researchers at UiO.

Outlook 2015

Bioinformatics at CEES will undoubtedly continue to grow in 2015. The following can already be said regarding the next year:

- all existing servers will be decommissioned
- the new servers (see above) will be made operational
- the AquaGenome project will generate much of its data in 2015, requiring significant computational resources for the analyses of these

Blindern, March 20th 2015
Lex Nederbragt, with help from many others.

The 'cod3' server. Photo: Lex Nederbragt