CEES bioinformatics – yearly report for 2013

Introduction

Bioinformatics – "an interdisciplinary field that develops and improves upon methods for storing, retrieving, organizing and analyzing biological data" (Wikipedia) is growing at CEES. 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 45 CEES members consider themselves bioinformaticians, a growth of more than 25% relative to 2012.

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 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 (some of their contribution we obtain as a paid-for service, see below). Finally, backup of the 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. As of December 31, 2012, the following computational infrastructure is 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 are 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 3 million CPU hours in 2013.
- we rent 30 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, 10 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

Investments in 2013

The life-time of the two oldest servers (cod1 and cod2, see above) was extended with one year by buying extra warranty.

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 with 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

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

- the project to generate an improved version of the Atlantic cod genome: both memory and CPU-intensive analyses, several TB of disk space
- several other genome sequencing projects (fish, bird)
- the RAD-seq platform at CEES (SNP detection and genotyping by sequencing): computations through the Stacks software on the cod1 server
- transcriptomics analysis pipelines
- 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 (45 subscribers), for general information exchange
- cees-hpc@bio.uio.no (25 subscriber) for specific information regarding shared Abel resources
  cod.nodes@.bio.uio.no (26 subscribers) 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 started a UiO wiki (https://wiki.uio.no/mn/bio/cees-bioinf). We are collecting articles dealing with the practicalities of using the resources at CEES, tips and trick, etc. The wiki is open to the world.

Courses

Continuing the success of previous courses (Unix, python and perl), in 2013 an internal courses was organized modeled upon the Software Carpentry (software-carpentry. org) Bootcamp curriculum. Furthermore, two CEES bio-informaticians (Karin Lagesen and Lex Nederbragt), as part of preparing for teaching a Software Carpentry Bootcamp, held a pre-bootcamp to try out teaching the material (four half days, spring 2013). From the website: 'Software Carpentry helps researchers be more productive by teaching them basic computing skills. We run boot camps 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.'

The actual bootcamp was held in July 2013, and was open for all researches. Around 25 people attended, including several CEES-bioinformaticians.

Outlook 2014

Bioinformatics at CEES is undoubtedly continuing to grow in 2014. The following can already be said regarding the next year:

- the life-time of the two oldest servers (cod1 and cod2) was extended until August 2013, but after this they most likely are going out of production soon. Losing these servers would mean a significant reduction in computational resources at CEES
- we are already running low on empty disk space and will have to increase with several TB to accommodate new datasets
- the AquaGenome project will generate much of its data in 2014, requiring significant computational resources for the analyses of these
- we will apply locally at the university for new computational infrastructure together with the Norwegian Sequencing Centre, the new 'endringsmiljø' Centre for Computational Inference in Evolutionary Life Sciences (CELS; Institute for Biosciences together with the Mathemical Institute, and the Institute of Informatics) and the Oslo node of ELIXIR.no. Preparations for the applications tarted already in December 2013.
- We will apply to norstore to increase our archiving storage space, including space on the tape-system for long-term storage. Much data that does not need to be accessed regularly will be moved over to norstore

Blindern, March **XX**th 2014
Lex Nederbragt, with help from many others.