Contribution ID: **53** Type: **Sessions contributions**

A virtual Biodiversity Lab

Tuesday, 20 May 2014 11:15 (15 minutes)

e-VirtualBiodiversityLab is a project with aGalaxy interface enabling researchers to run a series of tools classical in molecular biodiversity studies (alignment, phylogenies, and distances between sequences). A special emphasis has been put on taxonomic annotation from molecular data, i.e. annotate an unknown sequence with the label of the closest known reference in a database. In a first tool, we study the "shape"of the reference database (organized as clusters of taxa, or continuously varying pattern), and use this information to give an estimate of the quality of the taxonomic annotation. In the second tool, a file with a large number of unknown sequences coming from an environmental sample (like diversified eukaryotic community in a lake) is compared with a reference database, and an inventory is produced. The amount of sequences produced by Next Generation Sequencers requires work on scalability of these tools. The magnitude was a few thousands unknown sequences for one run some years ago, and current figures are closer to tens of millions. This task can be easily distributed, and EGI production grid is an ideal infrastructure for sharing the tool. A new layer has been added on the Galaxy server where the job can be launched on the grid, using Dirac as middleware. This permits the use of a e-lab for biodiversity on the grid. Next steps are to diversify and enrich those pipelines, and connect it with an IRODS implementation, for sharing data.

Wider impact and conclusions

The technological progresses in sequencing tools (NGS) have multiplied by several orders of magnitudes the amount of molecular data available for exploring biodiversity. These studies have been hampered by the cost of collecting data. This lock is vanishing with high throughput data production, and the way is open for revisiting and rescaling biodiversity studies. EGI infrastructure, providing computing and storing elements, is an ideal infrastructure for installing, running, and sharing codes and data (as with IRODS), and can help to gather the scientific community for wider applications. Solving scalability will necessitate better connections with applied mathematics (statistical modelling, machine learning, etc $\cdots$) and computer sciences. The developed Galaxy interface will enable access from any part of the world, including developing and emergent countries, and facilitate these exchanges.

URL(s) for further info

https://galaxy-pgtp.pierroton.inra.fr

Description of work

The work has consisted in
. collecting several classical public domain tools fro literature (alignment, phylogenies, etc.), and write the XML files for them to be launched from a galaxy server
. writing a code for sequence local alignment, with classical Smith-Waterman algorithm, which enables pairwise sequence comparisons (references and unknown) without heuristics, i.e. with exact calculation
. writing a code (in python) to distribute these codes a a cluster when necessary
. writing a code (in Python) for launching the code on EGI production grid using Dirac middleware
. design the Galaxy Interface for it to be user-friendly
. writing a manual explaining which tool does what, and how to select options.

The tools associate codes in C, Python, R and Perl, depending on the context and requirements. C codes for efficiency have been compiled to be callable as a libraries from python.

When a file of sequences is given, available tools permit
. to compute pairwise genetic distances
. to visualize the structure of the dataset by dimension reduction (Multidimensional Scaling)
. to visualize the structure of the dataset by representing it as a graph, with specimen being nodes, and an edge being drawn when the distance is under a given threshold. Visualization permits to select colors according to characters (like morphological species), enabling to evaluate the quality of inprints of species in variability of

molecular data.
. to have a taxonomic annotation for an unknown sequence, with a quality score taking into account the local shape of the reference, close to where the unknown sequence hits
. to produce an automatic inventory from an environmental sample of sequences, with number of sequences scaling with the recent sequencing technologies.

It is a project part of the Centre d'Etude de la Biodiversité Amazonienne, and it has been designed for being callable from French Guiana (or any other country). This project is open for collaboration.

Primary author: FRANC, Alain (CNRS)

Co-authors: Dr FRIGERIO, Jean-Marc (INRA); Dr CHAUMEIL, Philippe (INRA); Dr LAIZET, Yec'han (INRA)

Presenter: FRANC, Alain (CNRS)

Session Classification: Environmental science on grids and clouds

Track Classification: Success stories in using e-Infrastructures for research (Track Leaders: E. Katragkou, P. Castejon)