Multi-Cloud deployment for microbial genomes analysis

Jonathan LORENZO[1], Bryan BRANCOTTE[1], Thomas LACROIX[2] , Mohamed BEDRI[1], Jean-François GIBRAT[1] and Christophe BLANCHET[1]

[1] CNRS, UMS 3601 ; Institut Français de Bioinformatique, IFB-core, Avenue de la Terrasse, F-91190 Gif-sur-Yvette, France

[2] MaIAGE, INRA, Université Paris-Saclay, 78350 Jouy-en-Josas, France

Corresponding Author: jonathan.lorenzo@france-bioinformatique.fr, Thomas.lacroix@inra.fr

In the post-NGS area, sequencing bacterial genomes is very cheap (few hundreds €). Most of the time, users are no longer content to analyse a single genome; they want to compare large collections of related genomes (strains). This entails that biologists have to pay too much attention and dedicate their time to sequence the genomes, instead of thoroughly analysing the genomic data. Thus, this brings light to the increasing need for automating the annotation of bacterial genomes and carrying out efficient data mining.

In that context, IFB hub and the IFB-MIGALE platform developed a virtual environment (appliance), based on virtual machines, called "bacterial genomics" that aims to provide biologists and bioinformaticians access to suitable resources via the cloud. For example, Prokka [1] is a software tool for the rapid annotation of prokaryotic genomes. Insyght [2] developed by IFB-MIGALE is a tool for the visualization of the syntenies (local conservation of the gene order along the genomes) and the exploration of the landscape of both conserved and idiosyncratic genomic regions across multiple genomes. The platform automatically launches a set of bioinformatics tools (e.g. BLAST, HMMER, Prodigal…) to analyse the data and stores the results in a relational database (PostgreSQL). These tools use several public reference data collections. A web interface allows the user to browse the results. Setting up the platform requires solid skills in system administration since many bioinformatics tools with different dependences need to be installed as well as a relational database management system, a web server and servlet container, etc. Moreover, performing the analysis of a large number of genomes requires large computing resources and the use of parallel computing.

The goal is to deploy the “appliance” in one click over one or more cloud infrastructures. To achieve this, new features to automate deployment of complex application were added to the IFB’s cloud portal [3] through the connection to the SlipStream cloud broker [4]. Developed by SixSq, SlipStream is a multi-cloud application management platform. It automates the full application management lifecycle, within Infrastructure as a Service (IaaS) cloud infrastructures. Such complex application deployments can be done over several cloud infrastructures and provide scientists with high-level cloud features such as the dynamic allocation of a dedicated network for the isolation of the virtual machines, with the replication of the user data and with a direct link from the cloud portal to the Insyght web portal [5]. The appliance is available in the RAINBio catalogue of virtual images on the Biosphere web portal [6], and several tutorials on IFB bioinformatics cloud services usage are also available online on the main IFB website.

[1] Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014 Jul 15;30(14):2068-9. PMID:24642063

[2] Lacroix T., Loux V., Gendrault A., Hoebeke M., and Gibrat J.F. Insyght: navigating amongst abundant homologues, syntenies and gene functional annotations in bacteria, it's that symbol! Nucleic Acids Res. 2014 Dec 1; 42(21): e162. doi: 10.1093/nar/gku867 ; PMCID: PMC4245967

[3] https://biosphere.france-bioinformatique.fr/

[4] http://sixsq.com/products/slipstream/index.html

[5] https://cyclone.france-bioinformatique.fr/usecases/view/125

[6] https://biosphere.france-bioinformatique.fr/catalogue/appliance/19/