Experiences on metagenomic analysis using the EGEE Grid

I. Blanquer(1), V. Hernandez(1), G. Aparicio (1), M. Pignatelli(2), J. Tamames(2)

(1) Universidad Politécnica de Valencia - ITACA

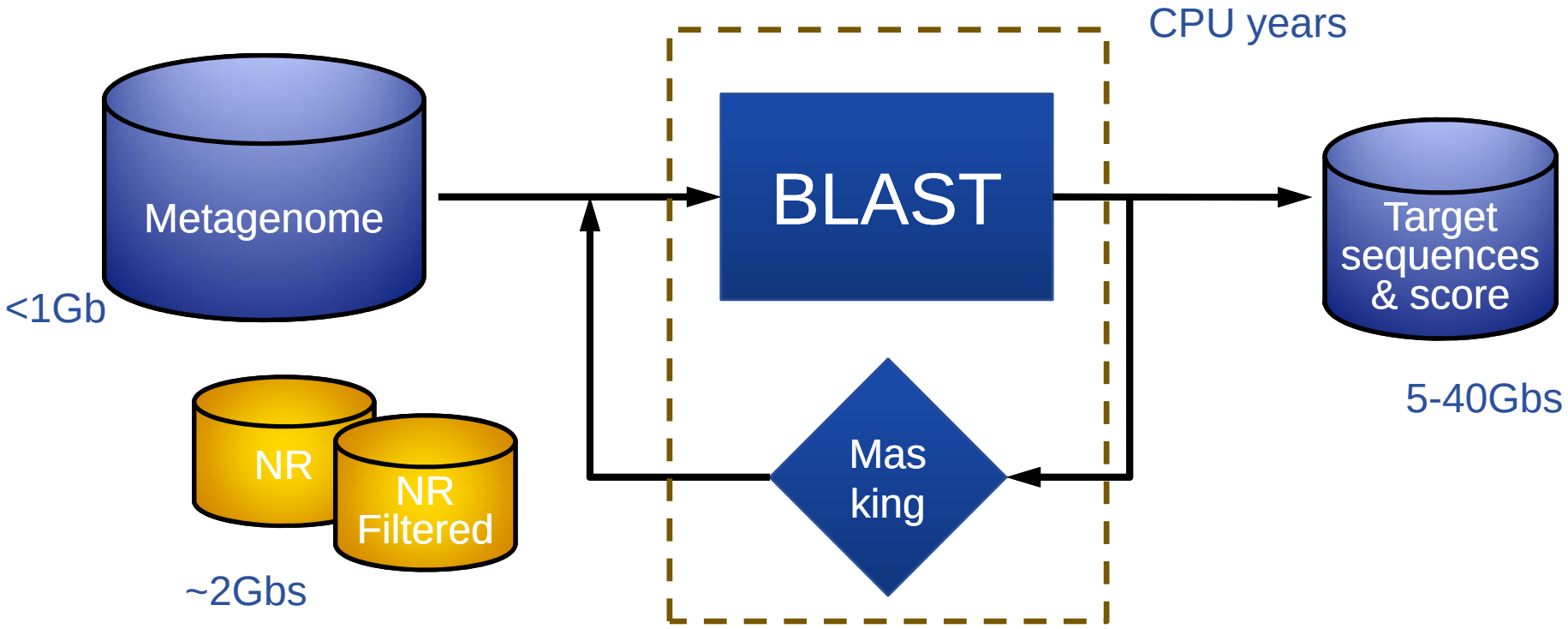
(2) Instituto Cavanilles de la Biodiversidad - Universidad de Valencia

Contents

- Metagenomic Analysis Execution Environment at the UPV
- Experiments Performed.
- Current Results and Conclusions.
- Work in Progress.

Metagenomic Analysis Execution Environment at the UPV

- **Objective: Perform the Alignment of Large Eukariotic and Prokariotic Metagenomic Samples.**
- **Procedure**



Steps of the Execution Script

- **Objective: Platform-Independent, Fault-Tolerant.**
  - Download BLAST from NCBI.
  - Configure and Install Locally.
  - Download and Reference Database from SE+Catalogue.
  - Execute Processing.
  - Copy and Register Output on SE+Catalogue.
  - Uninstall and Clean BLAST.

Steps of the Submission Schema

- **Preprocessing**
  - Download (NCBI), Filter, Reindex, Copy and Register the Reference Databases.
  - Split the Input Metagenome Sample.
  - Define the JDL Template and Create JDLs.
- **Execution**
  - Submit all Jobs.
  - Periodically Monitor the Status of the Execution.
    - Resubmit Failing Jobs or Jobs on Extremely Slow Resources.
- **Data Retrieval**
  - Copy, Check and Sort all Result Files from the Storage Elements.

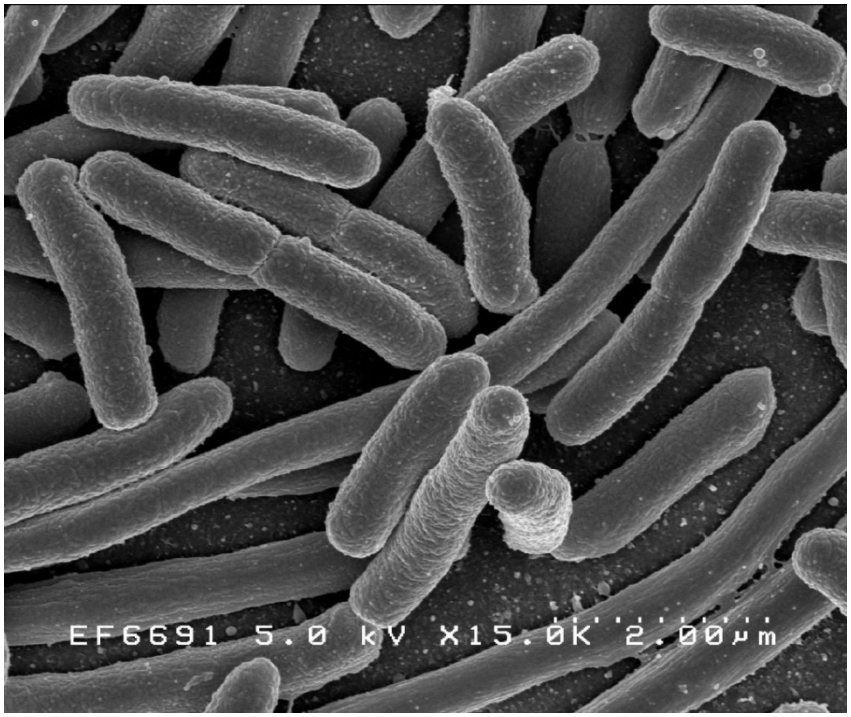
Case Studies - Biological

- **Farm Soil**
  - A Sample from a Nutrient-rich and Moderately Contaminated Soil Environment.
    - This Community is very Diverse and Complex.
    - Many yet Unknown Enzymes are Probably Present There.
- **Whale Fall**
  - Sample From a Whale Carcass.
    - They are Known to be a Nutrient-rich Environment in the Bottom of the Ocean.
    - A Heterogeneous Mixture of Bacteria Flourish There.

- **Sargasos's Sea**
  - Oceanic Samples Taken from Surface Waters.
    - They Represent the Diversity of Bacteria that Live Planktonically.

Case Studies – Medical

- **Gut Metagenomes**
  - Several Metagenomes of the Human Intestinal Microbiota.
  - A consortia of Bacteria that helps its host to Metabolize Many Nutrients that would be Indigestible Otherwise.
  - It is Involved in other Functions
    - Maturation and Modulation of the Immune Response of the Host.
    - Prevention of Infection by Bacterial Pathogens.

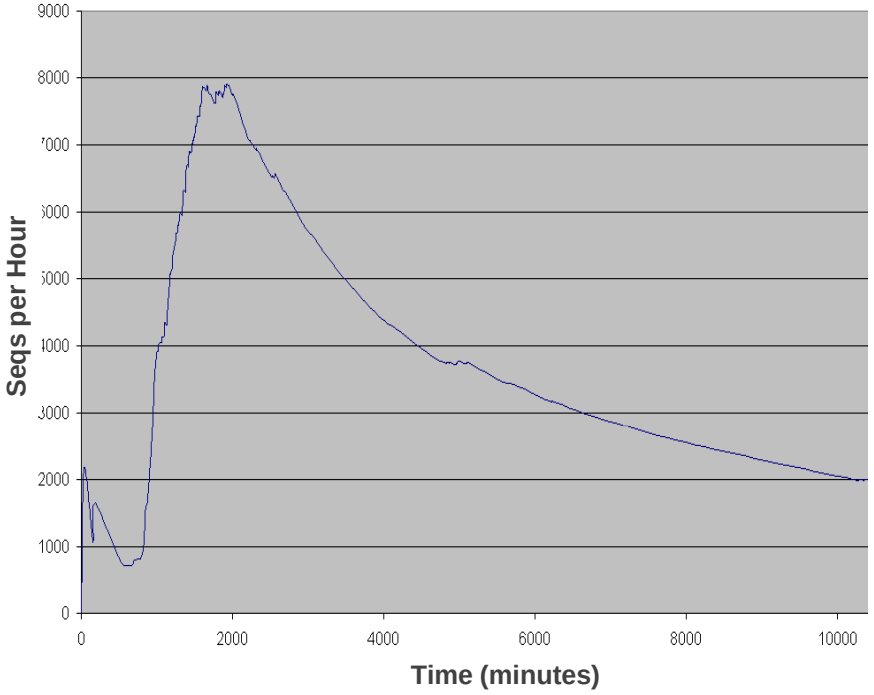


Resource Consume

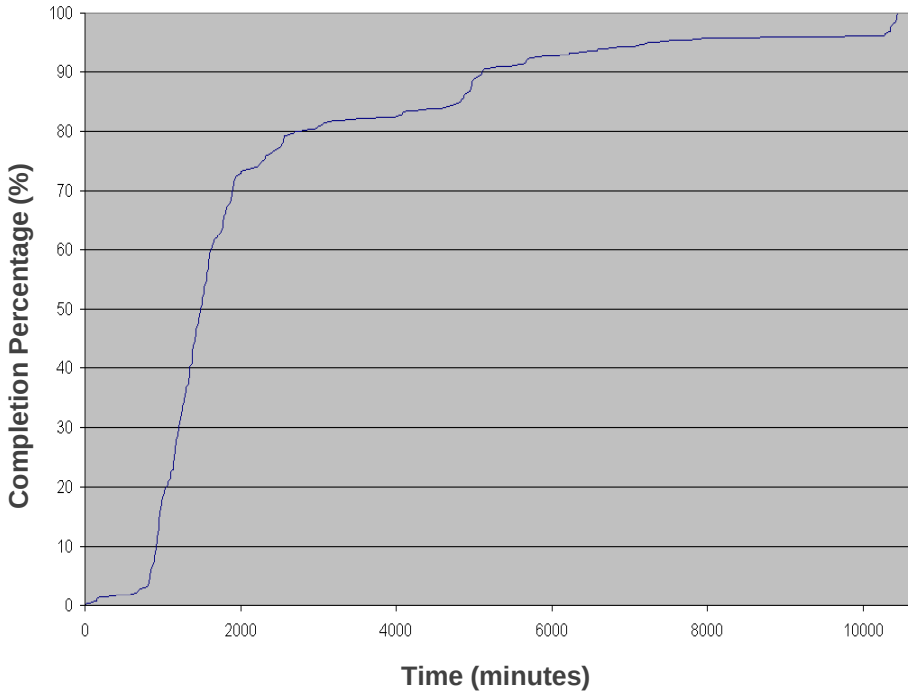
- The Different Experiments have Consumed more than 10 CPU Years with more than 20 thousand jobs, Producing More than 120 Gbytes of Effective Results.
- The Performance has Reached a Peak Performance of 8.000 Sequences per Hour (More than 120 Times Faster).
- The Load Balancing Still Requires Some Replication of Jobs To Ensure High Reliability and Quicker Response Time.

Metrics

Evolution of the Sequences Processed per Hour



Finished Jobs



Experiment

- **Objective**
  - Analysis of the Current Knowledge of the Universe of Sequences, and how This Knowledge has Evolved in the Past Years.
  - The possible Extent of Failures when Taxonomically Assigning ORFs to their Closest Relatives.
- **Methods**
  - Execution of BLASTX Searches for Several Metagenomes Against the Release 159 (April 2007) of GeneBank Non-redundant Database.
  - Restrict the list of the Homologues to those Already Present in a Particular Date, to Simulate the Results in that Time.

Conclusions

- **In the Experiment of Soil Farm the Results Showed that Many Assignments Have Changed, Even in Recent Times.**
- **The Rate of Change Does not Decrease in Recent Times, and Even it Increases for Most Cases.**
- **This Indicates that the Full Diversity of Theses Communities is Still not Well Described in the Current Databases.**

Ref: Miguel Pignatelli, Gabriel Aparicio, Ignacio Blanquer, Vicente Hernández, Andrés Moya and Javier Tamames, "Metagenomics reveals our incomplete knowledge of global diversity", Bioinformatics, ISSN 1367-4803, Oxford University Press 2008.

Further Work

- **Work is Continuing Through Four Ways:**
  - Execution of new Experiments.
    - Experiments of Virus Metagenomes have been Recently Started.
  - Migration of the Execution Environment to Different Problems
    - The Execution Environment is Being Generalised for Other Problems.
    - Currently we are Testing this Generalisation with the JAGS (Just Another Gibbs Sampler).
  - Extension of the Execution Pipeline with More Steps.
    - Creation of Phylogenetic Trees.
  - Refinement of the Resubmission System
    - A Trade-off Between Efficiency and Performance is Being Analysed to Increase the Number of Replicas in Failing Jobs.