Contribution ID: **61** Type: **not specified**

The Biovel Project: Robust phylogenetic workflows running on the GRID

Tuesday, 27 March 2012 12:00 (30 minutes)

Overview (For the conference guide)

Altered species distributions, the changing nature of ecosystems and increased risks of extinction all have an impact in important areas of societal concern. Biologists and environmental scientists are asked to provide decision support for managing the biodiversity component of our environment at multiple scales (genomic, organism, habitat, ecosystem, landscape) to prevent and mitigate such losses.
Biovel want to address this needs offering a series of robust and reliable web services that could be managed with the suite of tools of the myGRID project. The project will propose workflow for these services that will ensure best practice and efficiency of usage.
Within the first round of services produced by the project there are phylogenetic inference workflows.
These workflows will provide to end user the capabilities to execute application that could easily exploit several kind of resources, like: EGI grid infrastructure, local batch farm or dedicated servers.

Description of the Work

The workflow will start from a user supplied list biosequences, access an alignment webservice that implement HMMer3.0 align algorithm and using as guiding profile the correct PFAM profile chosen with the HMMer3.0 scan function and the query biosequences. Using a supplied user threshold the sites with lower posterior probability are filtered out. The alignment loaded in the workflow engine is them formatted to Nexus format. The model block is built following user supplied request, while the MCMCMC (Metropolis-coupled Markov Chains Monte Carlo) numerical integration options are in part from user specification in part are fixed to maximize MPI efficiency on the farm. The convergence of the numerical integration is controlled with the GeoKS software that estimate burn-in value and the realized convergence based on the tree parameter. The information is back supplied to MrBayes to produce summary statistics. The summary statistics is supplied back to the workflow. To control the fit of the model to the data a web service implement a posterior predictive test that use as input the sample from posterior distribution to simulate 200 data sets with the program Evolver (PAML suite) and compare the original data entropy with the distribution of simulated ones.
The workflow is build within Taverna Workflow Management System, each of the described steps are executed in a distributed computational environment like EGI grid infrastructure. This is possible because, we have build a REST-FUL web service, that exploits the usage of JST (Job Submission Tool) in order to submit and monitoring the jobs over the grid.
In this work we will show how the same web service build in Java and deployed over Tomcat server could be used to submit different applications and all procedures used to ensure the correct execution of the requested runs. We will also describe workflows provided to the final users and how they could help to use the grid infrastructure.

Conclusions

The solution described in this work will allow also the very end user to exploit the power of a computing grid infrastructure like EGI, without the complexity of learning a new interface. Indeed, the community of BioVel, as many others communities are used to have taverna as the only interface for their research.
Expressing the high level formalization of the algorithm in a workflow language allows scientists interest in algorithm but not expert in grid technology to improve and update the system, and in same time field scientist to use those services. In fact, using workflows, researchers could only focus the effort in scientific activities instead of learning complex procedures to execute their application, and once the workflow is developed all others researchers could re-use a part of it or the entire workflow to build something that is much more complicated.

Impact

Phylogenetic inference is a summary of the evolutionary history of a group of organisms. The topology summarize the relationships, while branch length summarize expected changes along a given section. So phylogeny is a basic tool to summarize biodiversity, categorize groups of organisms, and the study the impact of environmental change over the biodiversity. Unfortunately phylogenetic methods are both computationally intensive, and sensitive to misusage. This workflow will allow a broad adoption of best practice of phylogenetic inference in the current work of biodiversity scientist including ecologist and environmental scientist.
The usage of well configured workflow into taverna workflow manager, is the key advantage in this work, as it will allow the end user to manage the execution of complex algorithm with simple interaction, for example, configuring simple parameters such as input files and the option of the executable.
The use of JST help in the management of the jobs submitted to all computing infrastructure, and enable the Web Services to use all resources that are needed from the users.
In these workflows, indeed, the user could need different computing resources: grid EGI infrastructure, local batch facilities and dedicated servers. By means of those workflows and the use of JST, the end user could exploit all the resources in a transparent way.
To solve the problem of staging input and output, we choose a webdav server in order to keep the interaction between the users and the service as simple as possible. In fact in the webdav protocol the user could mount directly the remote server as a local file-system on his own personal computer, allowing a very easy transfer of single files or entire directory with a simple drag&drop.

Primary authors: Dr DONVITO, Giacinto (INFN-Bari); Dr NOTARANGELO, Pasquale (INFN-Bari); Dr VICARIO, Saverio (CNR-ITB)

Presenter: Dr DONVITO, Giacinto (INFN-Bari)

Session Classification: Community-tailored Services