The Biovel Project: Robust phylogenetic workflows running on the GRID

Saverio Vicario[1]✉, Bachir Balech[2], Giacinto Donvito[3], Pasquale Notarangelo[3], Graziano Pesole[2,4]

[1] Istituto di Tecnologie Biomediche, Consiglio Nazionale delle Ricerche, Bari, Italy
[2] Istituto di Biomebrane e Bioenergetica, Consiglio Nazionale delle Ricerche, Bari, Italy
[3] Istituto Nazionale di fisica Nucleare, Bari, Italy
[4] Dipartimento di Bioscienze, Biotecnologie e Scienze Farmacologiche,Università degli studi di Bari "Aldo Moro", Bari, Italy

Motivation and Objectives

Altered species distributions, the changing nature of ecosystems and increased risks of extinction all have an impact on important areas of social concern. Biologists and environmental scientists are asked to provide decision support for managing biodiversity components of our environment at multiple scales (genomic, organismal, habitat, ecosystem, landscape, etc...) to prevent and mitigate such losses. The BioVeL project (www.biovel.eu) is aspires to address these needs by offering a series of robust and reliable web services that could be managed with the tools suite of the myGRID project. The project proposes the building of workflows exploiting these services to ensure best practice and efficiency of use. These workflows provide the end users the capabilities to execute application easily accessible through several kind of resources such as EGI grid infrastructure, local batch farm or dedicated servers. Within the first round of services produced by the project, here we describe the phylogenetic inference workflows.

Phylogenetic inference is a summary of the evolutionary history of a group of organisms. The topology summarizes the relationships among the organisms, while branch lengths summarize the expected changes along a given section of them (Felsenstein, 2004). Therefore, phylogeny can be used as a basic tool to summarize biodiversity, categorize groups of organisms and study the impact of environmental change on biodiversity. Unfortunately, almost all phylogenetic methods are computationally intensive and sensitive to misuse (i.e bad model choice could cause high support for wrong answer). For that, this workflow offers an easy way to use phylogenetic services that will allow a broad adoption of best phylogenetic inference practices in the current work of biodiversity scientists including not only ecologists and environmental scientists (Honeycutt et al., 2010) but also medical doctors interested in studying patients' biome (Cho and Blaser, 2012; Delzenne et al., 2011). In particular, in the field of environmental sequencing processing biosequences within a phylogenetic context is a preliminary step for both taxonomic annotation and inferring evolutionary process from sequences within or across samples.

The usage of well designed workflow into Taverna workflow management system (Hull et al., 2006), is the key advantage of this work, as it will allow the end users to manage the execution of complex algorithms with simple interaction such as configuring simple parameters (input files and execution options), while the workflow will ensure the use of quality control steps and flag problematic inference at both the alignment level and then at the phylogenetic step itself.

It is important to note that, the implementation of the workflow within a workflow engine and editor publicly available, as in the myGRID suite of tools, allow two important practices to be implemented: 1) detailed peer review of the protocol implemented in a given work and 2) flexible update and/or modification of the workflow by the users without a specific coding capacities.

Methods

The workflow starts from a user defined list of biosequences (DNA/Amino Acids), access an alignment Web Service that implement HMMER3 align algorithm (Eddy, 2011) and uses, conditioned on the biosequences as queries, the correct PFAM as guiding profile chosen with 'HMMER3 scan' function. Using a supplied user threshold, DNA or Amino acids sites with lower posterior probability are filtered out. The alignment loaded in the workflow engine is then formatted to Nexus format. The MrBayes (Huelsenbeck et al., 2001; Altekar et al., 2004) model block is built following user supplied request, while the MCMCMC (Metropolis-coupled Markov Chains Monte Carlo) numerical integration options are in part specified by the user and in other part are fixed to maximize MPI efficiency on the farm system. MCMCMC numerical integration convergence is assessed by GeoKS (Battagliero et al., 2011) software that estimates burn-in value and the reached convergence based on the tree parameter.

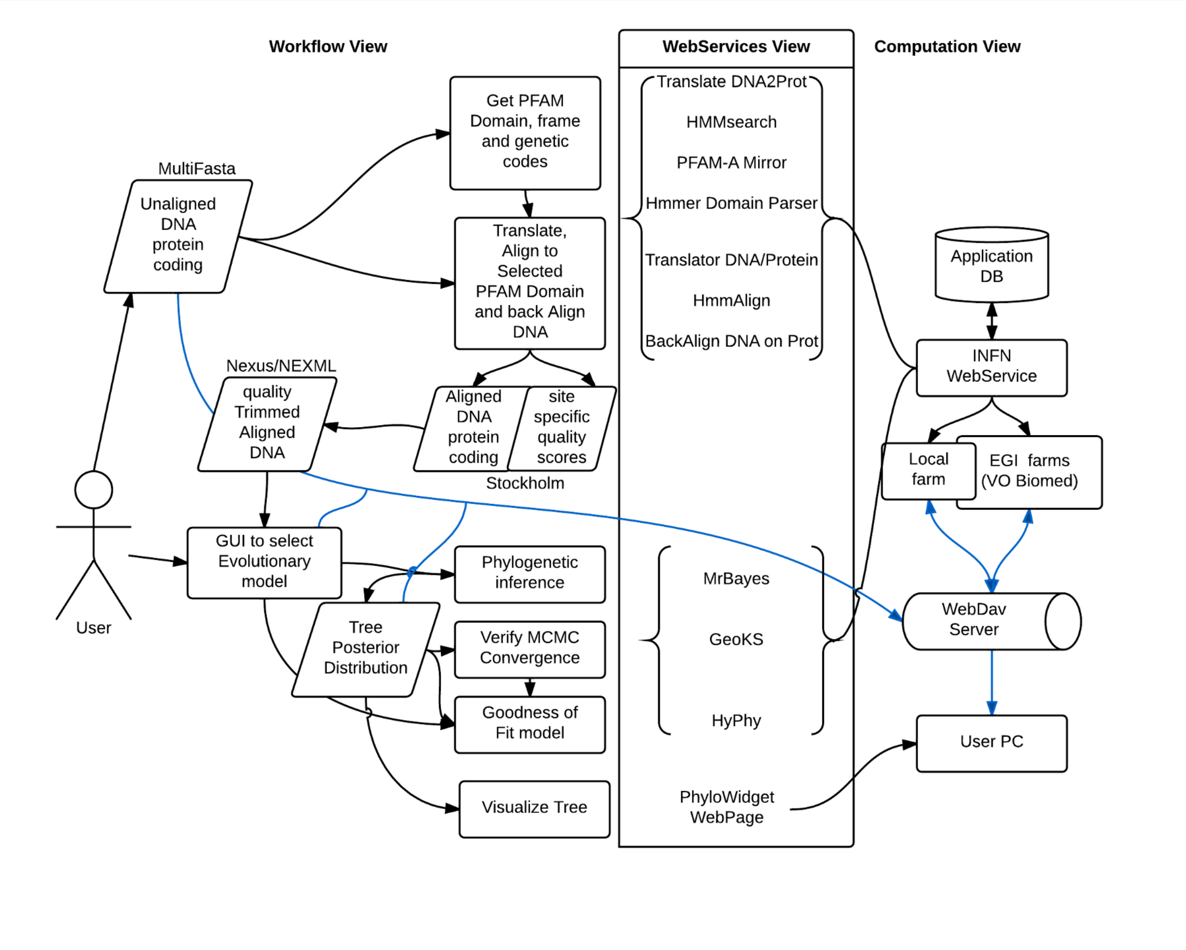

Figure 1. Schema Main Workflow showing underling web services and computational resources. Gray arrows indicates real data flow, black arrows logic and symbolic link. Table 1: Fragment of the information obtained with DiGSNP for Alzheimer disesase. A top rank of 5 genes is shown, and for each of them the top ranked 3 SNPs. Each SNPs is also associated with the region of the gene where it was found and the motif that generated its score.

The convergence information is back supplied to MrBayes to produce summary statistics submitted successively to the workflow. To control the molecular evolution model fit to the data, a web service implements a posterior predictive test within the software HyPhy (Pond et al., 2005) which uses as input the samples from the posterior distribution to simulate 200 data sets and compare the original data entropy with the distribution of simulated ones. The workflow is built 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 built a REST-FUL web service that exploits the usage of JST (Job Submission Tool) (DeSario et al. 2009; Tulipano et al. 2011) in order to submit and monitor the jobs over the grid. In this work we will show how the same web service built 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.

Results and Discussion

The use of JST helps in the management of jobs submitted to all computing infrastructure, and enables 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 and easy 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 using 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.

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 interested in setting up algorithm's parameters but not expert in grid technology to improve and update the system, and in same time non-expert scientists to use those services. In fact, using workflows, researchers could only focus the effort on scientific activities instead of learning complex procedures to execute their applications, and once the workflow is developed all others researchers can re-use a part of it or the entire workflow to build ad-hoc application according to their needs.

Acknowledgements

We would like to thanks FP7 funding that with the grant 283359 made possible this work

References

1. Altekar G, Dwarkadas S, Huelsenbeck JP, Ronquist F (2004) Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. Bioinformatics 20, 407-415.
2. Battagliero S, Puglia G, Vicario S, et al. (2011) An Efficient Algorithm for Approximating Geodesic Distances in Tree Space. Ieee-Acm Transactions on Computational Biology and Bioinformatics 8, 1196-1207.
3. Cho I, Blaser MJ (2012) APPLICATIONS OF NEXT-GENERATION SEQUENCING The human microbiome: at the interface of health and disease. Nature Reviews Genetics 13, 260-270.
4. De Sario G, Tulipano A, Donvito G, Maggi G, Gisel A (2009) High-throughput Grid computing for Life Sciences. In: M. Cannataro editor. Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare.
5. Delzenne NM, Neyrinck AM, Backhed F, Cani PD (2011) Targeting gut microbiota in obesity: effects of prebiotics and probiotics. Nature Reviews Endocrinology 7, 639-646.
6. Eddy SR (2011) Accelerated Profile HMM Searches. Plos Computational Biology 7.
7. Felsenstein J. 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Mass. 580pp.
8. Honeycutt LR, Hillis DM, Bickham JW (2010) Molecular approaches in natural resource conservation and management eds. DeWoody JA, Bickham JW, Michler CH, et al. Cambridge University Press.
9. Huelsenbeck JP, Ronquist F, Nielsen R, Bollback JP (2001) Evolution - Bayesian inference of phylogeny and its impact on evolutionary biology. Science 294, 2310-2314.
10. Hull D, Wolstencroft K, Stevens R, et al. (2006) Taverna: a tool for building and running workflows of services. Nucleic Acids Research 34, W729-W732.
11. Pond SLK, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21, 676-679.
12. Tulipano A., Marangi C., Angelini L., Donvito G., Cuscela G., Maggi G., & Gisel A. (2011). GRID distribution supports clustering validation of large mixed microarray data sets. EMBnet.Journal, 17 (1), pp. 18-25.