# Bayesian Phylogenetic Inference in the LIBI Grid platform

Vicario S(1), Mirto M(2), Aloisio G(2), Saccone C(3)

(1) ITB-CNR, Bari
(2) University of Salento, Lecce & SPACI Consortium
(3) ITB-CNR, Bari & Dipartimento di Biochimica e Biologia Molecolare
"E. Quagliariello", Università di Bari

## Motivation

The need to analyze larger data sets and the use of increasingly complex model of molecular evolution constrained in the last years the biologists interested in phylogenetic inference to also become interested, first in new statistical parameter optimization strategies, such as Metropolis-coupled Markov Chains Monte Carlo (MCMC), and subsequently in parallel and distributed computation methods to feed their increasing need for CPU. These are some of the reasons at the base of the LIBI (International Laboratory of Bioinformatics) project and its Grid Problem Solving Environment (PSE), built on top of EGEE, DEISA and SPACI infrastructures that allow the submission and monitoring of jobs mapped to complex experiments in Bioinformatics. In particular we ported MrBayes, the program for Bayesian phylogenetic inference, on the Grid through the PSE to allow, respectively, a highly parallelized (on DEISA and SPACI system) or highly distributed use on EGEE grid. The use of markovian integration as MCMC was accompanied by the adoption of the Bayesian statistics. This different framework allowed several advantages but also a main drawback: the estimates of the confidence error are highly dependent from the model. This requires not only an accurate choice among the most likely evolutionary models for each data set but also an estimate of the absolute fit of the model on the data. For this reason, we implemented in the LIBI PSE (Mirto et al. 2008), beside MrBayes, also a python script that implements the posterior predictive test, as proposed by Bolback (2002) and estimates the L statistics as proposed by Ibrahim et al. (2002). This to allows respectively an absolute evaluation of model and a comparative evaluation among several arbitrary models. To allow not experienced phylogenetists or user with several data sets to access the service, weimplemented a web interface.

## Methods

Within the LIBI PSE, we deployed both the parallel and the sequential version of MrBayes. In the first case, the program was used within SPACI and DEISA infrastructures, while the sequential version was used in the EGEE grid. The choice for the EGEE was guided by the fact, that although the parallel protocol (MPI) could be available, the minimal speed of communication among nodes is not guaranteed. To test the speedup and efficiency of the system we performed a series of phylogenetic inferences using several data sets for a total of 3523 sequences in 13 data sests, coming from a project of phylogenetic indexing of Barcode data base that our group is working on. The Barcode project is an international initiative whose goal is the development of standardized protocols for molecular species identification. This test set allowed us to tune the ratio between number of chains of the markovian integration and number of CPUs for our service, as it would be clearer in the results. The script for testing absolute and relative fit of evolutionary model is written in python and uses the Evolver program from the PAML package to produce the simulated data set necessary to obtain the posterior predictive distribution of data sets. The strength of the script and the novelty respect previous implementation of this test (MAPPS, Bolback, 2002) is the capacity to take in consideration a mixed model (i.e. that uses different submodels for each data partition) and to include also the L statistics of Ibrahim. The script understands the model description from the MrBayes output and accordingly instructs evolver in producing the simulated data set.

## Results

We have re-engineered and ported the MrBayes application, both sequential and parallel version on computational resources, available by using the LIBI Grid PSE. The web interface for MrBayes takes in input Nexus file or FASTA file (both as single or zipped directory). The FASTA files are automatically translated in a nexus format and the web interface guides the user in the description of a basic array of models for protein and DNA data and fill up the necessary MrBayes block, i.e. a set of options, in the newly formatted Nexus file. In case of multiple fasta files, mrbayes block are copied identical in all data set. Furthermore, the web interface, read each final input file two parameters relative to the markovian integration (nchains and nruns) in order propose two possible optimized solutions developed on our experience on the test data sets for the number of CPU to be used. Concluding, the advantages to use our service are: 1) to exploit a large data sets of data by using the grid resources; 2) to provide a user interface that allows composition of the input data needed to MrBayes; 3) to guide the user in the selection of best parameters in the computation; 4) to provide enhanced postprocessing analysis by using the combination of programs such as Evolver, in a transparent way to the user.

**Contact :** saverio.vicario@gmail.com