Emboss on gLite-Grid

Donvito G[(1)], Di Serio F[(2)], Maggi G[(1,3)], Gisel A[(4)]

[(1)]INFN Sezione di Bari, Bari
[(2)]Istituto di Virologia Vegetale, Sede Bari, CNR, Bari
[(3)] Dipartimento Interateneo di Fisica, Università degli Studi e Politecnico di Bari, Bari
[(4)]Istituto di Tecnologie Biomediche, Sede Bari, CNR, Bari

Motivation

EMBOSS, "The European Molecular Biology Open Software Suite" is a free Open Source software analysis package well established in the world-wide bioinformatics community. The analysis tools are mainly used in manual data analysis or in workflows via web services. The use is limited to a small set of data since the parallelization of using EMBOSS tools was not a major need up to now. However, running analysis on a genome-wide scale is out of the scope of using EMBOSS tools but most of the analysis tools are very suited for many different analysis types. The idea of using the EMBOSS package on the GRID environment is not new. A group within the EELA project developed GrEMBOSS (http://cimi.ccg.unam.mx/ccg-OrganicG/en/GrEMBOSS), a gridified version of EMBOSS. However, dealing with a large number of files GrEMBOSS demonstrated a weakness in data flow management. For this reason we followed a different approach to distribute the workload of EMBOSS analysis tools over the GRID. A test case demonstrated that the approach we were following was successful. Viroids are circular RNAs infecting plants They show compact secondary structures and are unable to code for any protein. Infectivity of these RNAs exclusively relies on their ability to interact directly with host factors (proteins and/or RNAs) and to redirect cellular machinery and biosynthetic pathways for their replication and spread in the host. Viroids accomplish this aim likely mimicking some host RNA structural property. Therefore, viroid RNAs may unveil structural motives with functional properties also contained in cellular RNAs. Bioinformatics approaches in viroid research are impaired by the fact that the complete genome of most natural viroid hosts is still unknown. To overcome this difficulty we decided to run a secondary structure analysis on sub-sequences of the whole plant sequence data set available in EMBL. We analysed as a first test 231’000 intron regions for the secondary structure of interest by using the vrnalfold algorithm (search for local folding patterns) from the EMBOSS/EMBASSY package.

Methods

Starting from a single file containing all the 231’000 sequences to be analyzed we produced several files with 1000 sequences in each.
These files were used to build the task in the JST task queue (http://webcms.ba.infn.it/cms-software/index.html/index.php/Main/JobSubmissionTool), then each task has 1000 sequences to analyze. As usual with JST, a daemon submits pilot jobs that pull from the DB Server the task to execute. As soon as they get the task an automatic procedure takes care of downloading emboss source code and compile them on the Worker Node. This will guarantee that the EMBOSS package will work on each machine on which the job lands regardless of it architecture and environment settings, it will also guarantee a high rate of success since the job is completely self-consistent. As the EMBOSS package is compiled, the job downloads the input file that contains 1000 sequences, and it splits this into 1000 files (one for each sequence). At this point vrnalfold is executed on all these files. At the end of each of this run, a perl script takes care to modify the output file in order to provide the information about the name of the input sequence. Are all 1000 runs terminated the output files are compressed in one single tar file and uploaded on a grid Storage Element (chosen between three different SE available).

Results

The initial test was executed on a file containing 231‘000 sequences using the whole EGEE glite infrastructure under BIOMED VO. Using the JST in order to manage the submission, it was quite easy to submit and control all the jobs. The complete run was performed in less than 2 hours and about 120 different Worker Node were used. The same run executed on a single CPU was estimated for 32 hours. During the submission we found several farms in which the EMBOSS set-up failed due to a lack of some basic library. This highlight the need to have such an advanced job management tool in order to try to fix the problem on the WN and in case to take care about the failure and the needed resubmission. The choice to pack the input and output files into blocks of 1000 files increased the efficiency in managing data over the grid, since here the

latency is one of the main problems. After the first successful run we are planning to execute another run in which we would optimize two aspects: the possibility to run more than one task of 1000 execution per job in order to improve the speed of the run, and the possibility to fix some of the compilation problem. The latter problem will be addressed, compiling and installing the needed library in the same procedure used with the EMBOSS package. From the bioinformatics point of view we were successful to find the secondary RNA structure of interest in a very small set of sequences but from their background information, some predictions could be interesting in the context of viroid-plant interaction.

Contact : andreas.gisel@ba.itb.cnr.it