Bioinformatics, 1(1), 1–2
Advance Access Publication Date: 8 April 2022
Applications Note

Phylogenetics

ModelTest-NG: Parallelized selection of evolutionary models for DNA and proteins

Pedro Sánchez[1,*]

[1]Master in Bioinformatics for Health Sciences, University of A Coruña, 15071 A Coruña, Spain

*To whom correspondence should be addressed.

Associate Editor: Jorge González

Abstract

Motivation: We are in a context in which next generation sequencing technologies provide data sets of considerable size. This factor, in phylogenetic analyses, implies the need for advanced computational resources and constant modifications in the development of tools that allow the scientific community to work efficiently.
Results: We have worked with the high-performance computing (HPC) version of ModelTest-NG, a tool that combines the features of jModelTest and ProtTest, used for years for projects in various fields. ModelTest-NG is remarkably faster than its predecessors, adding new models, parameters, heuristics, and technical optimizations for parallel execution.
Availability: ModelTest-NG source code, full documentation and tutorials are freely available at https://github.com/ddarriba/modeltest.
Contact: p.sanchezg@udc.es

1 Introduction

Next-generation sequencing (NGS) technologies have generated great advances in recent years, allowing scientists to work on more ambitious and larger-scale genomics projects. However, the main problem is determined by the large amount of data with which it works. In phylogenetic analyses, multiple sequence alignments comprise hundreds or thousands of sequences, implying a transition to large computing infrastructures. In this way, high performance computing (HPC) provides support and performance for these types of processes that consume abundant resources and time.

It should be noted that scientists specialized in this type of research demand tools that implement a large number of evolutionary models (Sullivan & Joyce, 2005). This is fundamentally due to the fact that the selection of the best evolutionary model for an alignment of DNA or protein sequences is a very important step in phylogenetic analysis. In addition, the tools must be flexible in the parameters available for the algorithms, such as topology or model heterogeneity rates, and must be updated to include new proposals and technical modifications for correct performance. In this sense, ModelTest-NG represents a novel tool that is based on the progress made with jModelTest and ProtTest (Abascal et al., 2007; Posada, 2008), reinforcing the selection of models in divergent and long sequences. In this way, ModelTest-NG allows working with molecular evolution models and adjusting numerous parameters so that the researcher can acquire a realistic view of the evolutionary process for the multiple sequence alignment with which he works.

2 Methods

The multithreaded version of ModelTest-NG based on PThreads was installed to work on shared memory systems. The level of parallelism that is produced is of coarse-grained type, where initially, a set of models is evaluated and provides optimized parameters, so that, in the subsequent parallel evaluation, the threads possess this information, with less synchronization for the remaining process. To do this, first of all, the tool was downloaded in compressed *.tar* format from the link: https://github.com/ddarriba/modeltest. Subsequently, using the scp client, the tool was transferred to the user account provided for Finis Terrae II at the Supercomputing Center of Galicia (CESGA). The tool was uncompressed using *-tar xvf* and the GNU Bison dependency was loaded, which consists of a parser that prepares the text in a tabular format for parsing. In the same way, the Flex dependency was loaded, which provides support for the tools used to search patterns in sequences. Once this step was done, the tool was compiled with CMake 3.12.4, designed to generate software packages.

To verify a first test of the tool, an example set in FASTA format with 430 16S rRNA sequences from various worm species was executed in the login node. As a result, a time of 2 minutes was reached for the calculation and generation of the best evolutionary models in this multiple alignment.

With ModelTest-NG, the researcher can work with multiple alignments of DNA or protein sequences in FASTA or PHYLIP format, adjust the basis of tree construction (maximum likelihood by default), set base or amino acid frequencies in function of the project, and propose the tool that acts as a basis in the generation of all possible evolutionary models. In this work, a standard evaluation of the 88 evolutionary models present in the tool was chosen, taking the default parameters, which allows an optimization of the evaluation process of evolutionary models and the generation of the best tree considered.

In order to benchmark the performance of ModelTest-NG in a Finis Terrae II node, 3 datasets were chosen (**Table 1**) from the HIV databases (http://www.hiv.lanl.gov/), corresponding to sequences that code for the Pol, Gag and Env proteins, key in different phases of the HIV cycle and isolated from patients in the United States.

Table 1. Summary of datasets used in the performance evaluation.

Name	Number of sequences	Length	Sequential runtime (hh:mm:ss)
Pol	4125	3540	4:00:58
Gag	6574	2126	1:46:29
Env	8075	4130	3:32:16

Sequential runtime is the time required to calculate the likelihood scores and generate the best models using the sequential version (i.e a single thread).

For each of them, the running time was computed for the estimation of the best models and reconstruction of the best phylogenetic tree with 2,4,8,12,16 and 24 threads. Numerous scripts were used to launch these configurations in CESGA, which has a SLURM system.

3 Results

The multithreaded version of ModelTest-NG executed on a node of the Finis Terrae II (24 cores: two 12-core Intel Xeon Haswell 2680v3 processors) gives the results shown in **Table 2**. In general, with the dataset Pol a drastic reduction in execution time is achieved on 2 cores compared to 1 core. In the case of the Gag dataset, a less remarkable reduction is achieved than in the case of Pol. A similar pattern is produced for the Env dataset, where we can see that with 24 cores, a significantly higher execution time is achieved than the other datasets. This result could be associated with the length of the sequences and the presence of numerous gaps, which reduces the evaluation of those complex evolutionary models.

Table 2. Execution times (hh:mm:ss) with different configurations on a single node.

Threads	Pol	Gag	Env
1	04:00:58	01:46:29	03:32:16
2	01:47:25	01:00:21	02:09:18
4	00:57:45	00:35:24	01:14:33
8	00:32:38	00:22:05	00:47:20
12	00:26:08	00:19:13	00:41:49
16	00:17:23	00:17:34	00:36:40
24	00:15:18	00:12:07	00:32:50

The analysis of this shared memory architecture with 24 cores was complemented through speedup (**Fig. 1**), referring to $Speedup = T_{seq} / T_n$, where T_{seq} is the running time of the sequential execution of ModelTest-NG, and T_n the time measured when using n cores. In general, the scalability was linear with up to 12 threads, but scaled well with 16 threads in the case of the Gag and Env datasets and with 24 threads with the Pol dataset.

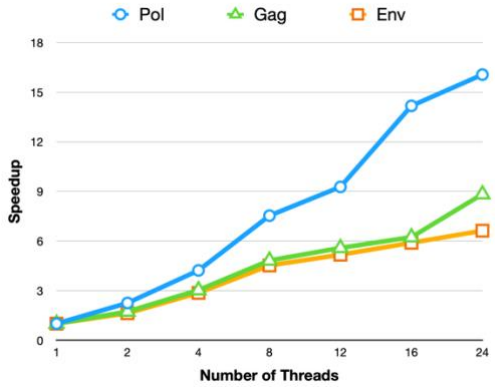

Fig 1. Speed-ups obtained with the multithreaded version of ModelTest-NG according to the numbers of threads used in a 24-core shared memory node (two 12-core Intel Xeon Haswell E5-2680v3 processors) with 128 GB memory.

With the previous figure we can observe the previously mentioned trend with the execution times, where the speedup is lower in the Gag and Env datasets. It should be noted that with Model-Test-NG, the greater the number of threads, there is a loss in the optimized workload distribution per thread, although less when compared to the predecessor jModelTest (Darriba et al., 2014). Due to this issue, we can see that the scalability tends to a saturation with 24 threads.

Despite these performance issues, in terms of speed, ModelTest-NG has been shown to be 100 times faster than jModelTest (Darriba et al., 2019). In this context, it could be equivalent to a reduction in execution time from about 8 days to 1 hour. In this way, it is evident that the parallelized version of the tool is a support to speed up the evaluation phase of evolutionary models in ambitious research on phylogeny.

References

Abascal,F., Zardoya, R., & Posada, D. (2007) ProtTest: selection of best-fit models of protein evolution. *Bioinformatics*, *21*(9), 1104-1105.

Darriba, D., Taboada, G. L., Doallo, R., & Posada, D. (2014). High-performance computing selection of models of DNA substitution for multicore clusters. *The International Journal of High Performance Computing Applications*, *28*(1), 112-125.

Darriba, D., Posada, D., Kozlov, A. M., Stamatakis, A., Morel, B., & Flouri, T. (2019). ModelTest-NG: A New and Scalable Tool for the Selection of DNA and Protein Evolutionary Models. *Molecular Biology and Evolution*, *37*(1), 291–294.

Posada, D. (2008). jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution*, *25*(7), 1253-1256.

Sullivan, J. & Joyce, P. (2005). Model selection in phylogenetics. *Annual Review of Ecology Evolution and Systematics*, *36*(1), 445-466.