Phylogenetic analysis of ITS data from Endophytic fungi using Massive Parallel Bayesian Tree Inference with Exabayes

Análisis Filogenético de Secuencias ITS Provenientes de Hongos Endófitos Utilizando Inferencia Bayesiana Paralela de Árboles con Exabayes

Maripaz Montero-Vargas[1], Jean Carlo Umaña-Jiménez[2], Efraín Escudero-Leiva[3] , Priscila Chaverri-Echandi[4]

Montero-Vargas, M; Umaña-Jiménez, J; Escudero-Leiva, E; Chaverri-Echandi, P. Phylogenetic analysis of ITS data from Endophytic fungi using Massive Parallel Bayesian Tree Inference with Exabayes. *Tecnología en Marcha*. Edición especial 2020. 6th Latin America High Performance Computing Conference (CARLA). Pág 74-79.

https://doi.org/10.18845/tm.v33i5.5079

1 Biologist. Centro Nacional de Computación Avanzada (CNCA), CeNAT-CONARE, Costa Rica. Email: mmontero@cenat.ac.cr .
https://orcid.org/0000-0002-6562-4231
2 Computing Engineer. Centro Nacional de Computación Avanzada (CNCA), CeNAT-CONARE, Costa Rica. Email: jumana@cenat.ac.cr .
https://orcid.org/0000-0003-0857-6007
3 Biologist. Centro Nacional de Innovaciones Biotecnológicas (CENIBiot), CeNAT-CONARE, Costa Rica & Centro de Investigaciones en Productos Naturales (CIPRONA), Universidad de Costa Rica, Costa Rica.
https://orcid.org/0000-0003-4440-4296
4 Associate Professor. Universidad de Costa Rica, Escuela de Biología & Centro de Investigaciones en Productos Naturales (CIPRONA), Costa Rica.
https://orcid.org/0000-0002-8486-6033

Keywords

Fungi; ITS; Exabayes; Phylogenetics; Parallelization; Biodiversity.

Abstract

Ecological studies of fungal communities have been favored thanks to the emergence and improvement of independent culture techniques that use the ITS region as a molecular marker. This has allowed a more accurate identification compared to traditional culture-dependent methods.

Next-generation sequencing techniques have increased the amount of data available for the understanding of endophytic fungal communities. An important part of this process is the phylogenetic inference to decipher how the different taxa are related and interact, however, this may become one of the bioinformatic analysis that demands more time.

In response to this, the bioinformatics along with high-performance computing offer solutions to accelerate and make more efficient the tools available for data processing through the implementation of supercomputers and the parallelization of tools

In this study we carried out the processing of ITS sequences to then use the parallelization of Exabayes, software specialized in the analysis and creation of phylogenetic trees.

Thanks to the use of this technique, it was possible to reduce the running time of Exabayes from more than 400 hours to 6 hours, which demonstrates the benefits of the use of high-performance computing platforms.

Palabras clave

Hongos; ITS; Exabayes; Filogenética; Paralelización; Biodiversidad.

Resumen

Los estudios ecológicos de las comunidades fúngicas se han visto favorecidos gracias a la aparición y mejora de técnicas independientes de cultivo que utilizan la región ITS como marcador molecular. Esto ha permitido una identificación más precisa en comparación con los métodos tradicionales dependientes de la cultura.

Las técnicas de secuenciación de próxima generación han aumentado la cantidad de datos disponibles para la comprensión de las comunidades de hongos endofíticos. Una parte importante de este proceso es la inferencia filogenética para descifrar cómo se relacionan e interactúan los diferentes taxones, sin embargo, este puede convertirse en uno de los análisis bioinformáticos que exige más tiempo.

En respuesta a esto, la bioinformática junto con la informática de alto rendimiento ofrecen soluciones para acelerar y hacer más eficientes las herramientas disponibles para el procesamiento de datos a través de la implementación de supercomputadoras y la paralelización de herramientas.

En este estudio llevamos a cabo el procesamiento de secuencias ITS para luego utilizar la paralelización de Exabayes, software especializado en el análisis y creación de árboles filogenéticos.

Gracias al uso de esta técnica, fue posible reducir el tiempo de ejecución de Exabayes de más de 400 horas a 6 horas, lo que demuestra los beneficios del uso de plataformas informáticas de alto rendimiento.

Introduction

Endophyte fungi have been found in almost all vascular plant species examined to date and they are considered important components of fungal biodiversity [1]. They can have profound effects on their host physiology, influence multitrophic networks and entire ecosystems [2]. Fungi and particularly endophytes are a very promising source of novel biologically active compounds [3] so studying them may help to implement new techniques for new biotechnological applications.

An essential part of ecological studies of endophytic fungi is the taxonomic classification and phylogenetic inference [4], [5]. Fungi taxonomy is certainly complex, even now, what was believed to be one species can be in fact be an assemblage of productively isolated lineages. In mycology, species were defined based on the morphology of asexual and sexual reproductive structures, unfortunately, the number of characters is limited generating insecurity in the identification of species [6].

The easiest way to identify endophytic fungi is through molecular methods. ITS region is accepted as a barcode for fungi because of the higher amplification success rate for many fungal groups [7] [8]. Thanks to the next generation sequences the amount of data obtained in the ecological studies has increased exponentially through the years [9], [5], which in turn implies a challenge for the processing and analysis of the information.

Recently is common to join computational efforts, the development of bioinformatics tools and the and specialized databases to make the analysis of the large sets of biological data more fast and efficient [10], [11]. High performance and high throughput computing technologies permit that the processing of such datasets could be automated to accelerate bioinformatics processes [12].

One of the most time-consuming analyses is the Bayesian Tree inference [13]. This is why a critical step in the ecological analysis is to use parallel tools that fulfill this task.

Exabayes is an is a software package that computes Bayesian tree posterior probability using the Markov Chain Monte Carlo sampling approach. This tool applies commonly used evolutionary models and can handle big data sets efficiently. An important aspect of this tool is that it allows the use of Message-passing Interface (MPI) to parallelize the analysis using a computer cluster so that the only limit for the analysis of large data sets is the memory held by the cluster [14].

Exabayes works by performing a calculation of the posterior probability of sampling trees. This tool can divide its analysis on multiple independent runs which in turn analyze multiple chains responsible for sampling the parameters for stochastic simulation to obtain a sample from the posterior distribution of trees [15], [14].

In this paper, we implement the analysis of ITS sequences of endophytic fungi of coffee plants to achieve the reconstruction of their phylogeny through the parallelization of exabayes with the objective of improving the performance of this tool using a computational cluster. These techniques may help ecologist or micologist to process their data in a efficient way, improving the use if computational resources and generating more accurate results.

Methodology and results

For the following experiments, a set of sequence data in Fastq format was used, which were obtained from endophyte fungi extracted from Costa Rica coffee plants. Then a quality control of the sequences was carried out using the FastQC [16] tool for the identification of poor quality bases. All bioinformatic processing was carried out in the Kabre cluster of the National Center of High Technology in San Jose, Costa Rica

The preprocessing of the sequences was made with the Seqtk tool [17], using two of its functions. First, the Trimfq function was used to remove the poor quality bases at both ends of the sequences, identified with a Phred score of less than 20. Then the Seq function was used to change the format of the Fastq sequences to FASTA.

Then duplicate sequences were removed with the USEARCH [18] tool with which a single sequence multifasta file was obtained with 331 sequences of 1331 each. The identification of the Operational Taxonomic Units (OTUs) was then carried out using the UNITE database [10] through the Blast tool of the National Center for Biotechnology Information [19].

For the phylogenetic analysis, an alignment of the sequences was performed using the MUSCLE tool [20]. The resulting file in phylip format is used as input for the Tree inference with Exabayes.

Initially, the analyzes in Exabayes were carried out using nodes, composed of Intel Xeon Phi KNL nodes processors, each one with 64 cores @ 1.3 GHz and 96 GB (architecture A). Historically. A sequential test was performed using 256 ranks and dividing the analysis into four runs and these in turn into four chains using four swaps between chains. This test lasted 447 hours.

To increase efficiency, parallelization of runs and chains was implemented using MPI the parameters described in the documentation: number of runs executed in parallel (-R), number of chains per run executed in parallel (-C) and number of swap attempts between chains per generation.

From these experiments, the best performance was obtained using 64 ranks, and dividing the analysis into 2 runs with 2 strings each. This same distribution was used for parallel runs and parallel chains, with 4 swaps per generation. This test lasted 20:29:01 (hh: mm: ss) reaching statistical significance with = 0.0131916 (table 1).

These tests were also performed with a different architecture (architecture B) using nodes c composed of nodes with Intel (R) Xeon (R) CPUs E3-1225 v5 @ 3.30GHz and 16GB RAM. In the same way, as with the previous tests, the use of the parameters was compared in parallel and better and statistically significant performance (asdsf = 0.0199349) was obtained using 4 ranks and dividing the analysis into 2 runs and 2 chains. It was also specified that the two runs and the two strings will be executed in parallel. With this the best result was obtained with a run time with 6:25:31 (hh: mm: ss) (table 1).

Analysis

As shown in the previous results, Exabayes efficiency improved substantially by balancing the number of runs and chains in which the analysis is divided with the number of runs and chains executed in parallel.

In this way, the better performance of the traditional test tool is observed to test 1. Thus, the number of ranks can be better distributed among the groups and achieve better performance.

With these tests, it was also found that by decreasing swapping among coupled chains, the distribution of resources becomes more efficient by decreasing communication between groups (of runs and chains that are being executed in parallel.

On the other hand, when comparing the performance of Exabayes between the different architectures, there is a significant improvement in the tests carried out on nodes B with respect to nodes A. This indicates that it is more advantageous for Exabayes, to use fewer CPUs but with more clock speed (3.3 GHz), than using a node that has a lot of processors with a very low clock speed (1.3 GHz).

Table 1. Results obtained from the evaluation of efficiency in the use of Exabayes in parallel.

Architecture	Number of Ranks	Runs in parallel	Chains in parallel	Runs	Chains	swaps	Wall time (hh:mm:ss)	asdfs
A	256	NA	NA	4	4	4	447:36:22	0.014756
A	64	2	2	2	2	2	9:44:05	0.0156383
B	4	2	2	2	2	2	6:25:31	0.0199349

Conclusions

The use of the parallel modality of the Exabayes tool is shown as an excellent alternative to improve the efficiency of tree inference with the Bayesian method, allowing better run times with statistically significant precision. In this matter, is important to balance the number of runs and the number of chains runs in parallel so that a homogeneous distribution of the workload of the groups in the different ranks is achieved

We recommend making an evaluation not only of the input data to be used during the analyzes but also on the available computational resources that can give Exabayes more efficiency for phylogenetic inference.

For future studies, it is recommended to scale the data set to a more complex one in terms of number of taxa and length of the sequence. A next step would be to apply data level parallelism in alignments with multiple partitions, so that the scalability of the balance of the parameters used in this experiment can be tested.

References

[1] J. Rodríguez, J. Elissetche and S. Valenzuela, “Tree Endophytes and Wood Biodegradation”. *Endophytes of Forest Trees*, pp.81-93, 2011.

[2] M. Unterseher, “Diversity of fungal endophytes in temperate forest trees”. In Endophytes of forest trees. Springer, Dordrecht. pp. 31-46, 2011.

[3] T. Larsen, J. Smedsgaard, K. Nielsen, M. Hansen and J. Frisvad, “Phenotypic taxonomy and metabolite profiling in microbial drug discovery”. *Natural Product Reports*, vol. 22, no. 6, pp. 672, 2005.

[4] J. Fouquier, et al. “Ghost-tree: creating hybrid-gene phylogenetic trees for diversity analyses”. *Microbiome*, vol. 4, no. 1, 2016.

[5] S. Tibpromma, “Identification of endophytic fungi from leaves of Pandanaceae based on their morphotypes and DNA sequence data from southern Thailand”. *MycoKeys*, vol. 33, pp.25-67, 2018.

[6] C. Grünig, V. Queloz and T. Sieber, “Structure of Diversity in Dark Septate Endophytes: From Species to Genes”. *Endophytes of Forest Trees*, pp.3-30, 2011.

[7] C. Schoch, C., et al., “Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi”. Proceedings of the *National Academy of Sciences*, vol. 109, no. 16, pp. 6241-6246, 2012.

[8] U. Kõljalg, et al. (2013). “Towards a unified paradigm for sequence-based identification of fungi”. *Molecular Ecology*, vol. 22, no. 21, pp.5271-5277, 2013.

[9] J. Zoll, E. Snelders, P. Verweij and W. Melchers, “Next-Generation Sequencing in the Mycology Lab”. *Current Fungal Infection Reports*, vol.10, no. 2, pp. 37-42, 2016.

[10] K. Abarenkov, *et al.*, “The UNITE database for molecular identification of fungi - recent updates and future perspectives”. *New Phytologist*, vol. 186, no. 2, pp.281-285, 2010.

[11] C. Wurzbacher, “Introducing ribosomal tandem repeat barcoding for fungi”. *Molecular Ecology Resources*, vol. 19, no. 1, pp.118-127, 2018.

[12] A. Welivita, I. Perera, D. Meedeniya, A. Wickramarachchi and V. Mallawaarachchi, "Managing Complex Workflows in Bioinformatics: An Interactive Toolkit With GPU Acceleration". *IEEE Transactions on NanoBioscience*, vol. 17, no. 3, pp.199-208, 2018.

[13] G. Altekar, S. Dwarkadas, J. Huelsenbeck and F. Ronquist, "Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference". *Bioinformatics*, vol. 20, no. 3, pp.407-415, 2004.

[14] A. Aberer, K. Kobert and A. Stamatakis, (2019). ExaBayes: Massively Parallel Bayesian Tree Inference for the Whole-Genome Era.

[15] B. Rannala and Z. Yang, "Probability Distribution of Molecular Evolutionary Trees: A New Method of Phylogenetic Inference". *Journal of Molecular Evolution*, vol. 43, no. 3, pp. 304-311, 1996.

[16] S. Andrews, (2010). FastQC: a quality control tool for high throughput sequence data.

[17] H. Li, (2012). seqtk Toolkit for processing sequences in FASTA/Q formats.

[18] R. Edgar, "Search and clustering orders of magnitude faster than BLAST". *Bioinformatics*, vol. 26, no. 19, pp. 2460-2461, 2010.

[19] G. Boratyn, et al., "BLAST: a more efficient report with usability improvements". *Nucleic Acids Research*, vol. 41, no. W1, pp. W29-W33, 2013.

[20] R. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput". *Nucleic Acids Research*, vol. 32, no. 5, pp.1792-1797, 2004.