Molecular Phylogenetic Analysis: Design and Implementation of Scalable and Reliable Algorithms and Verification of Phylogenetic Properties

Jorge Álvarez-Jarreta, Gregorio de Miguel Casado, Elvira Mayordomo

Grupo de Ingeniería de Sistemas de Eventos Discretos (GISED),
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: *jorgeal@unizar.es*

Abstract

Since the creation of the bioinformatics field, and even more since the creation of the so called *next-generation sequencing*, the relevance of computer methods and technologies has significantly increased. We here present several contributions we have published as solution for different problems involved in the study of the evolution process.

Background

Phylogenetics is the study of the evolution process for one or more species. Many different methods have been developed and published to solve the different problems involved in these studies. Once the dataset of biological sequences has been selected, the first step is to align them [1,2]. The difference in lengths can appear due to sequencing errors (*digitalizing* the biological sample), mutations (insertions or deletions of one or more sites along the sequence) or because the researcher also wants to include fragments of the same genetic region that were used in other experiments. The second stage is the phylogenetic inference, where an evolutionary tree is estimated throught different methods or evolution models [3,4].

Several software systems have been designed and implemented to cope with the whole phylogenetic inference process aforementioned. Some have been developed for specific biological data, like ZARAMIT [5], and others are of general purpose, even large case scenarios, like SATé or DACTAL [6,7].

The phylogenetic analysis comes right after a phylogentic tree has been inferred. The study of the conservation index is one of the most common analyses, where different levels of molecular distance (phylogenies involving a set of sequences of the same species, of closely-related species or any different set of species) allows to calculate how well preserved has been a set of substrings of the input biological sequences through time (and, therefore, evolution) [8]. This results can be used to determine the relevance of different parts of the genome of different species, being the most conservative sites the most suitable to not being spreaded through time if any mutation affects them.

Materials and Methods

Phylogenetic inference

PhyloFlow is the first phylogenetic inference system fully customizable and automatic for novel and expert users [9]. It has been designed with workflow techniques and implemented under *Condor+DAGMan*. The the systems mentioned in the background lack of adaptation to the user needs fixing the tools used as well as their parameterization. Moreover, many of these systems have not been designed to handle large-case scenarios, making them not suitable for input datasets of more than a few thousand sequences.

Due to the high time and economic cost of inferring a phylogenetic tree (more than a week for tens of thousand sequences), the update processes should happen at most every three or four months. Meanwhile, the new sequences that might provide relevant information for biological studies are hold until the next reconstruction. We proposed PHYSER [10] as a new method to detect possible sequencing errors and as an update process for phylogenetic trees in the time between full-tree updates.

Phylogenetic analysis

The first stage of a phylogenetic tree analysis usually involves the visualization of the tree. Several tools have been proposed for this process, but many of them provide an intractable interface when the input phylogenetic tree has more than a few hundred sequences. Thus, we created

PhyloViewer [11], a visualization tool intended for extense phylogenetic trees.

In collaboration with F. Merino-Casallo, we studied different methods to calculate and study the conservation index for large input alignments. As a result, we published a new software tool to measure the conservation index under different methods and techniques using parallelization techniques [12]. Furthermore, we studied the impact of different alignment tools and different parameterizations on the conservation score for the same datasets.

Conclusion

We have here presented several contributions on the phylogenetics fields, designed and implemented to solve different problems involved in the inference and analysis processes of evolutionary trees.

As future work, we will expand the methods available and we aim to improve their efficiency and throughput for very large-case scenarios.

REFERENCES

[1]. EDGAR, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. In: *Nucleic Acids Research*. 2004, 32(5), 1792-1797. Available from: doi:10.1093/nar/gkh340.

[2]. Katoh, K., MISAWA, K., KUMA, K., and MIYATA, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. In: *Nucleic Acids Research*. 2002, 30(14), 3059-3066. Available from: doi:10.1093/nar/gkf436.

[3]. PRICE, M.N., DEHAL, P.S. and ARKIN, A.P. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. In: *Molecular Biology and Evolution*. 2009, 26, 1641-1650. Available from: doi:10.1093/molbev/msp077.

[4]. STAMATAKIS, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. In: *Bioinformatics*. 2006, 22(21), 2688-2690.

[5]. BLANCO, R., MAYORDOMO, E., MONTOYA, J., and RUIZ-PESINI, E. Rebooting the human mitocondrial phylogeny: an automated and scalable methodology with expert knowledge. In: *BMC Bioinformatics*. 2011, 12, 174.

[6]. LIU, K., RAGHAVAN, S., NELESEN, S., LINDER, C.R., and WARNOW, T. Rapid and Accurate Large-Scale Coestimation of Sequence Alignments and Phylogenetic Trees. In: *Science*. 2009, 324, 1561.

[7]. NELESEN, S., LIU, K., WANG, L.-S., LINDER, C.R. and WARNOW, T. DACTAL: divide-and-conquer trees (almost) without alignments. In: *Bioinformatics*. 2012, 28, i274-i282.

[8]. PEI, J. and GRISHIN, N.V. AL2CO: calculation of positional conservation in a protein sequence alignment. In: *Bioinformatics*. 2001, 17(8), 700-712.

[9]. ÁLVAREZ-JARRETA, J., DE MIGUEL CASADO, G., and MAYORDOMO, E. PhyloFlow: A Fully Customizable and Automatic Workflow for Phylogenetic Reconstruction. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2014, pp. 1-7.

[10]. ÁLVAREZ-JARRETA, J., MAYORDOMO, E. and RUIZ-PESINI, E. PHYSER: An Algorithm to Detect Sequencing Errors from Phylogenetic Information. In*: 6th International Conference on Practical Applications of Computational Biology & Bioinformatics (PACBB)*. 2012, pp. 105-112.

[11]. ÁLVAREZ-JARRETA, J. and DE MIGUEL CASADO, G. PhyloViewer: A Phylogenetic Tree Viewer for Extense Phylogenies. In: *13th European Conference on Computational Biology (ECCB)*. 2014.

[12]. MERINO-CASALLO, F., ÁLVAREZ-JARRETA, J., and MAYORDOMO, E. Conservation in Mitochondrial DNA: Parallelized Estimation and Alignment Influence. In: *IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 2015.