REPRODUCIBLE AND SCALABLE DATA ANALYSIS ON HIGH PERFORMANCE COMPUTING

Emrah AKKOYUN[1], Nurdan Kuru[1], Onur Dereli[1], Aylin Bircan[1], Öznur Taştan[1] and Ogün ADEBALİ[1,2]

[1] Faculty of Engineering and Natural Sciences, Sabancı University
Orta Mahalle, 34956, Tuzla, İstanbul, Turkey
[2] TÜBİTAK Research Institute for Fundamental Sciences, 41470 Gebze, Turkey
phone: + (216)5687043, email: emrahkyn@gmail.com, ogun.adebali@sabanciuniv.edu

ABSTRACT

We recently presented the PHACT tool (PHylogeny-Aware Computation of Tolerance) for assessing amino acid substitutions that achieved superior predictive performance compared to widely adapted tools [1]. PHACT scores alterations not only using the frequency of the alterations in the multiple sequence alignment (MSA) as most common tools do - but also uses the gene-based phylogenetic trees. PHACT's inputs include the MSA of the protein, the phylogenetic tree estimated on that MSA and the probability distribution of amino acids at each ancestral node estimated from the tree. To assess the predictive performance of PHACT. We performed various experiments over a dataset that include 20,546 proteins and 61,662 variants.

In theory, analyzing a protein takes a single CPU day using eight cores, thus, the amount of computation is 192 CPU hours. The overall computation time to finish the analyses for the whole dataset (20.546 proteins) is 3.94M CPU hours. Using a single and powerful computer with 64 cores takes around seven years, and 512GB of memory is not practical. We completed the analyses within four months by using a High-Performance Computing (HPC) cluster.

Performing an extensive reproducible and scalable data analysis for multiple proteins with various parameters on an HPC is not straightforward [3]. For example, 50 proteins with ten parameters and ten consecutive tasks mean 5000 independent jobs must be executed successfully. Each job (task) has different characteristics; some are CPU, and others are memory-intensive jobs. Some jobs are completed within hours, while others take days or weeks. On the other hand, HPC is a complex environment; hundreds of servers running together and obtaining a failure is not an exception; thus, managing such a large number of jobs is not easy. A workflow tool, where an analysis definition is determined by a set of rules and a set of output files from a set of input files is obtained, is a must. In addition to being scalable, being reproducible is also a critical requirement that the same results can be obtained by other researchers anytime [4]. All the tools, software used during the analyses, input files, the computational facilities could be defined in a text file so all environments could be deployed without additional efforts. To satisfy all these requirements, we used a Snakemake workflow with a conda package manager due to its human-readable, Python-based language, portability, integration with a conda package manager, automatic deployment, and ability to specific software dependencies.

PHACT framework specifies rules in Snakefile. Rules decompose the workflow into small steps such as finding homologs of each query sequence (PSI-BLAST), performing multiple sequence alignment (MAFFT), or generating a maximum-likelihood phylogenetic tree (RAXML-NG, FASTTREE). Each rule has its model parameters, which can be set via a single configuration file (config/config.yml). A dry-run parameter can be used to check if the workflow is adequately defined and to estimate the amount of calculation remaining. It summarizes the number of total jobs (rule) performed and sets of input and output files used and created, respectively. For 2 query files, as given in Fig.1, 29 jobs will be executed. In addition, to allow workload running on a local computer with a limited number of query IDs, the PHACT framework is designed to analyze a bulk of query IDs in parallel using HPC. Most HPC clusters have a scheduler that handles the workload on compute nodes. Users must prepare a bash script and submit it to the cluster to interact with a scheduler. Snakemake has the functionality to perform all these efforts automatically.

Within this work, a valuable dataset that contains MSAs and phylogenic trees, which amounts to more than 1M files and a 1.6TByte size, was created and shared with other researchers. All details such as documentation, scripts, tools, environments, and input proteins can be found on our GitHub page [5] and all results are published on our FTP server [6].

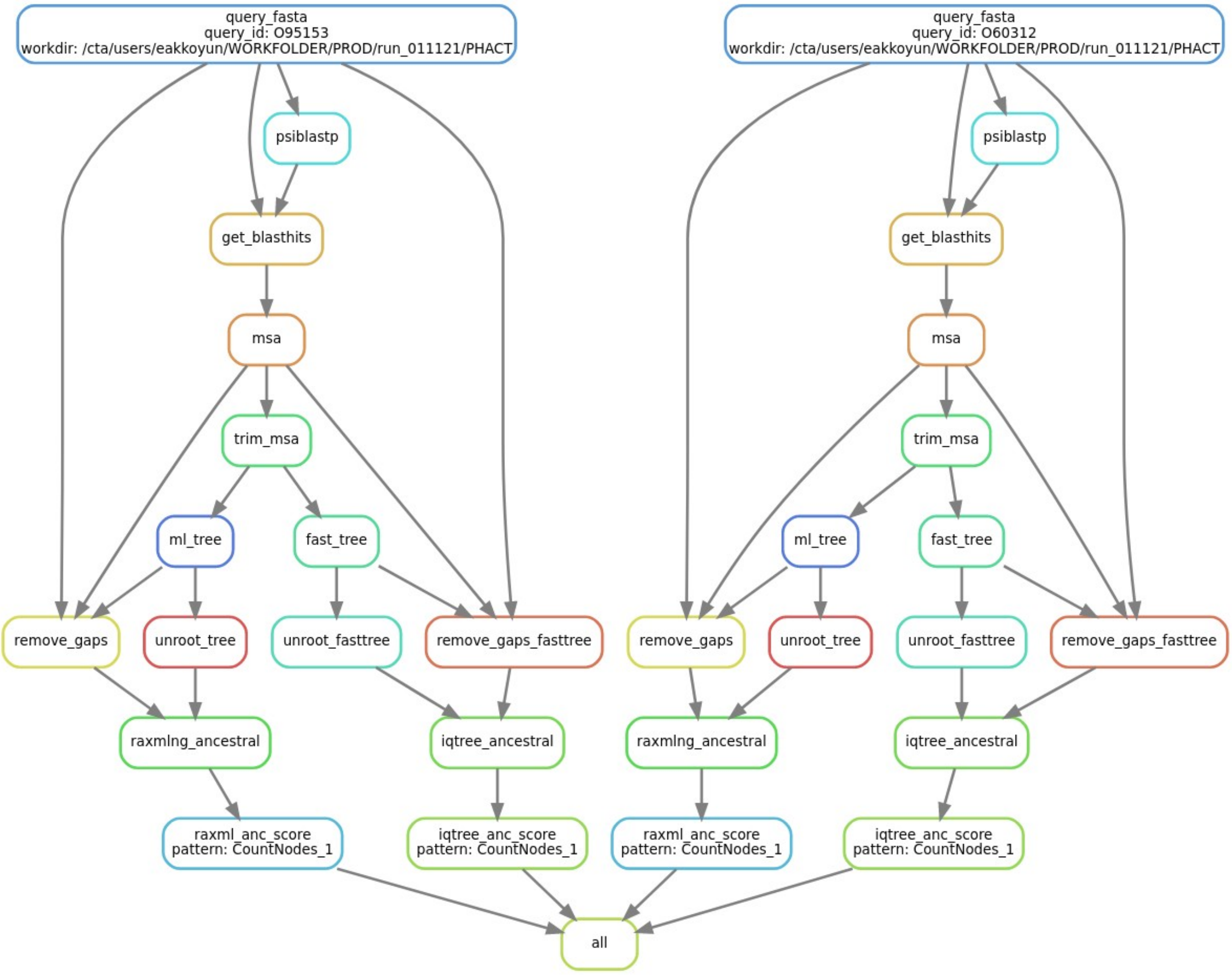

Figure 1. The workflow of PHACT for given queries is an example of predicting the effect of missense mutations.

References

[1] Kuru, Nurdan, et al. "PHACT: Phylogeny-aware computing of tolerance for missense mutations." Molecular Biology and Evolution 39.6 (2022): msac114.
[2] Köster, Johannes, and Sven Rahmann. "Snakemake—a scalable bioinformatics workflow engine." Bioinformatics 28.19 (2012): 2520-2522.
[3] Wratten, Laura, Andreas Wilm, and Jonathan Göke. "Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers." Nature methods 18.10 (2021): 1161-1168.
[4] Mölder, Felix, et al. "Sustainable data analysis with Snakemake." F1000Research 10 (2021).
[5] https://github.com/CompGenomeLab/PHACT)
[6] https://phact.sabanciuniv.edu/pubs/kuru_mbe_2022/