Evolution in a nutshell: HTC as a tool to resume million of years into tenths of seconds.

Javier F. Tabima, M. Sc.

Laboratory of Mycology and Phytopathology,

Group of Computational and Evolutionary Biology,

Department of Biological Sciences,

Universidad de los Andes.

Bogotá, Colombia.

Introduction

As time happens by, organisms changes in a way that they can deal with everyday issues in order to survive to the next generation (Darwin, 1859). As simple as it may seem, it takes millions of years (in most cases) to develop changes that creates new species and the high level of diversity that leads to today. As so, evolutionary reconstructions such as phylogenetic trees try to recreate the contexts of morphological or molecular changes in a scale of time which can relate to the real time of divergence of each species and ancestor(Omland et al., 2008).

In order to reconstruct phylogenetic histories, basic information about the identity of the species are requested, such as Nucleic acids, proteins and morphological traits. Additionally, ecological and dating samples are needed in order to be able to reconstruct changes in a time scale relation(Drummond et al., 2002). Of course, advanced mathematics are needed to understand the probabilistic of evolution (such as Likelihood and Bayesian probabilistic approaches), and this mathematical performance requires several machines that can develop such complicated tasks, whereas a human researcher will take several months in developing all the phylogenetic history of 10 taxa (a possibility to find the most likely tree from a dataset of ≈ 35 Million trees), a single-node machine will last several seconds.

As so, computational resources have been fundamental for the understanding of evolution, whereas different software uses several algorithms to increase computational efficiency and mathematical complexity as they reduce computational time so biologist can work in the results. But lately, in the new era of genomics, the evolutionary relationships that were reconstructed using a gene or two have become obsolete, as whole-genome reconstruction are state of the art. This represents a new problem for bioinformaticians everywhere, as classic resources (from one to eight cpus per node) are not enough to deal with such an amount of data (the lowest cpus for a decent reconstruction using a likelihood approach are 16 cores in a average speed of 3GHz).

As computer resources are expensive for the third world countries, we propose a method to increase efficiency without the need of a lot of resources, but several machines with one or to processors, in order to develop a HTC approach to reconstruct the phylogenetic history of large datasets.

Phylogenetic reconstructions occur in a three dimensional space of likelihood (or posterior probabilities (PP)) against number of generations or “steps”. The number of generation is translated as computing cycles, as in each step is a new likelihood/PP value. Additionally, the coordinate where an X step finds a Y likelihood/PP is called a seed number. This seed numbers represents a newly-found evolutionary tree. For ease of understanding, I’ve compressed this interactions in a two-dimensional way as can be seen in Figure 1.

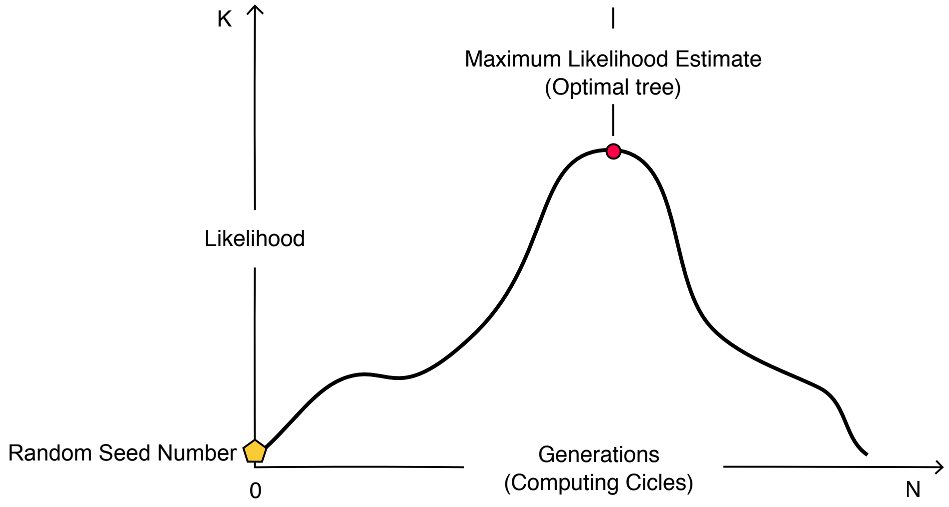
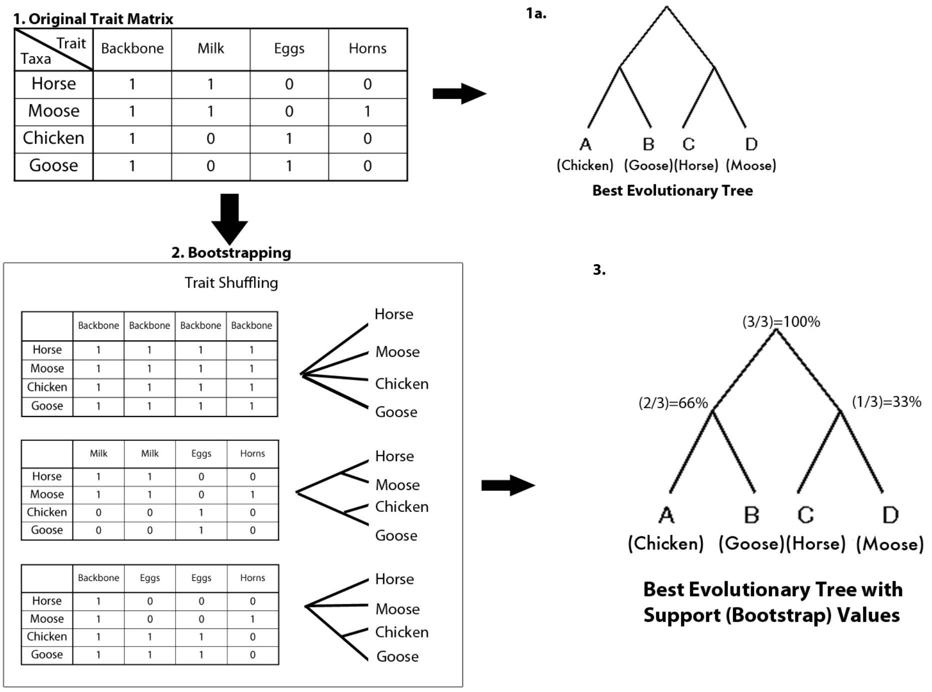


Figure 1. Sampling space for likelihood values (2D figure). Seed number is the coordinate whereas a step or generation has certain likelihood value.

Phylogenetic software requires: Sequence Data, Model of evolution data (defined as the probabilities of changes between nucleotides/protein against time, using present time representations) and Random Seed number, whereas the reconstruction will start. But additionally, one of the main requirements is the bootstrap number (For ease of understanding I will not use the Bayesian Monte Carlo Markov Chain concept for this part).

Bootstrapping, in a context of evolutionary trees, is defined as the statistical process that permits the non-parametric sampling of different hypothesis (e.g. Data Shuffling, Taxa Shuffling, ect.) to inform about the support of different phylogenies(Guindon et al., 2010; Müller, 2005). As so, bootstrap values are also known as support values, and are vital for the statistical assurance of a clade or taxa group (Figure 2 explains better the bootstrapping method for evolutionary trees). In this order, evolutionary reconstructions requires a standardized bootstrapping number: When bootstrap replicates are few, the phylogenetic reconstruction will have underestimation of the clades reconstruction; When bootstrapping is high, overestimation of the support values can lead to non-evolutionary relationships with significant bootstrap value (>50% support of the branch containing the clade). Recommended values for bootstrapping are between 1000-2000 replicates per reconstruction(Müller, 2005).



1. Original Trait Matrix

Trait / Taxa	Backbone	Milk	Eggs	Horns
Horse	1	1	0	0
Moose	1	1	0	1
Chicken	1	0	1	0
Goose	1	0	1	0

2. Bootstrapping

Trait Shuffling

	Backbone	Backbone	Backbone	Backbone
Horse	1	1	1	1
Moose	1	1	1	1
Chicken	1	1	1	1
Goose	1	1	1	1

	Milk	Milk	Eggs	Horns
Horse	1	1	0	0
Moose	1	1	0	1
Chicken	0	0	1	0
Goose	0	0	1	0

	Backbone	Eggs	Eggs	Horns
Horse	1	0	0	0
Moose	1	0	0	1
Chicken	1	1	1	0
Goose	1	1	1	0

Figure 2. Bootstrap explained for Phylogenetic Reconstruction. 1. The original trait table with each of the evolutionary traits and species. 1a. best tree found by that matrix. 2. bootstrapping, represented as shuffling of the traits, whereas each one has its best tree. 3. mapping bootstrap values over the best tree found. It can be seen that the results of the bootstraps shows that the horse is not so well grouped with the moose, because the traits shuffled did not show any concordance between them and the original tree. (Tree modified from http://bit.ly/puFgsl).

Bootstrapping is a limitant factor in both reconstruction and computational time (Stamatakis et al., 2008). If a likelihood estimation is 4sec. long, one hundred bootstrap replications will be approximately 400sec., a number that grows if the number of taxa grows. As so, phylogenetic analyses using whole genomes (Phylogenomics) are painfully slow, taking days and even months, and requires a huge amount of resources (Running 25 taxa, 970 proteins, ≈200kaminoacids with 1000 bootstrap replicates in a 8 core CPU took 1 month, 4 days and 5 hours to complete).

One way to avoid the bootstrap complication is to run several reconstructions one after another, but this method does not accommodate to a parallel reconstruction of the trees, and takes almost the same amount of time if the bootstrap replicate standardization is not found in a intermediate number. I propose to break the bootstrapping values intro 100-200 and using them to start at different random seeds, whereas the dimensional space of likelihood will be fully scanned in less time and phylogenetic reconstructions will be faster and more accurate.

Methods

Using the DAGman tool, several parent jobs containing a script with different random seed numbers will be ran at the same time, using the same amount of processors and memory requirements (one or two cores, 3GHz and 2GB RAM). Hard disk requirements will not be as problematic as each of the resulting files is less than 30Kb with 100 trees in it, and phylogenomic files with 20 species are up to 50Kb of size. We could run up to 10 jobs of 100 bootstrap replicates each. Afterwards, a tree-concatenations script will be ran, to determine the likelihood values contained in the treefile for each tree (1000 trees in total) and the likelihood values will be compared in the R statistical suite, using Akaike's criteria, Bayesian criteria and a hierarchical Likelihood ratio test to determine the best likelihood value in all trees, and this tree will be selected as the optimal tree reconstruction for the group of organisms studied (Figure 3).

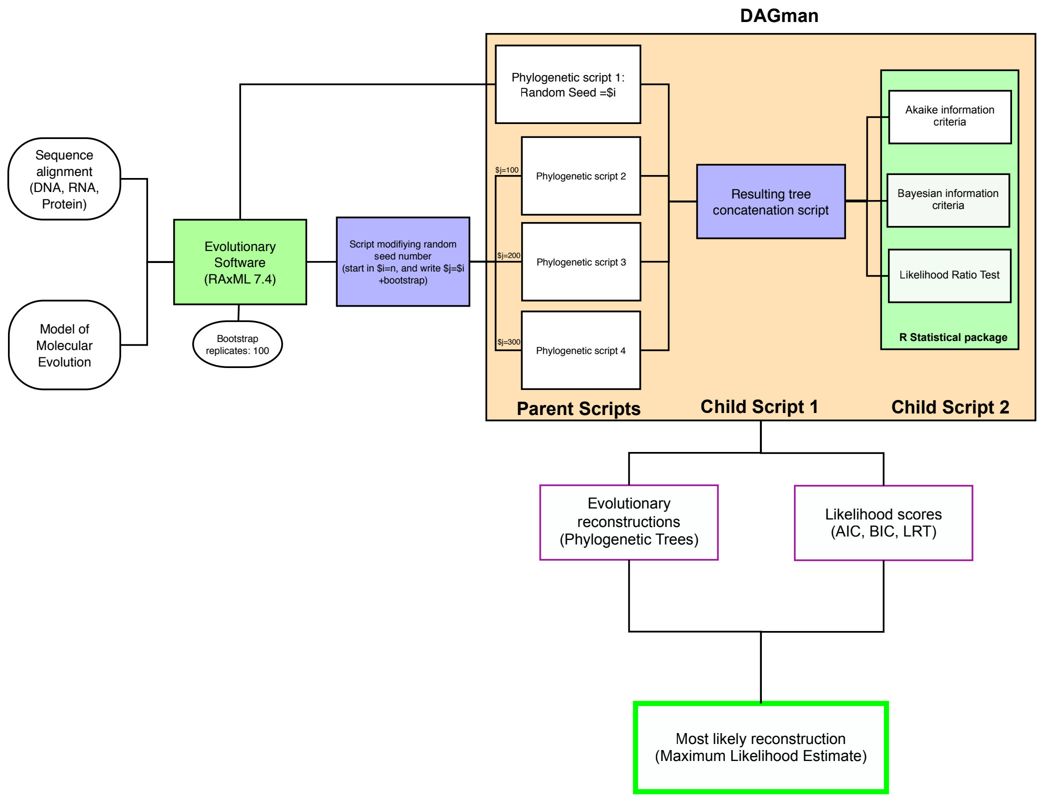

DAGman
Sequence alignment (DNA, RNA, Protein)
Model of Molecular Evolution
Evolutionary Software (RAxML 7.4)
Bootstrap replicates: 100
Script modifiying random seed number (start in $i=n, and write $j=$i +bootstrap)
Phylogenetic script 1: Random Seed =$i
$j=100
Phylogenetic script 2
$j=200
Phylogenetic script 3
$j=300
Phylogenetic script 4
Resulting tree concatenation script
Akaike information criteria
Bayesian information criteria
Likelihood Ratio Test
R Statistical package
Parent Scripts
Child Script 1
Child Script 2
Evolutionary reconstructions (Phylogenetic Trees)
Likelihood scores (AIC, BIC, LRT)
Most likely reconstruction (Maximum Likelihood Estimate)

Figure 3. Pipeline for the development of the method. Green: Software to be used. Yellow: Pipeline administrator. Blue: Scripts written by the author. In purple stroke is the results from the pipeline and in green stroke the final and processed result.

This approach will be valid in all probabilistic phylogenetic methods, as Bayesian inference also uses Random Seed Numbers and the bootstrap values are substituted by MCMC generations, whereas MCMC tries to sample the entire space of evolutionary hypothesis (Beerli, 2006; Huelsenbeck and Ronquist, 2005; Nylander et al., 2004).

Insights

We will implement this method in a 5 node cluster, 24 Core/32Gb each, in order to determine the phylogenomic history of the Oomycetes (A group containing plant pathogens (*Phytophthora infestans, Pythium spp.*) and human/animal pathogens (*Saprolegnia parasitica*). At this moment, we are not able to run the jobs since a transition from PBS and SGE to Condor is being made in this cluster.

Acknowledgments

I would like to thank the Universidad de los Andes, Bogotá, Colombia and the Center of investigation for Coffee (CENICAFE), Colombia. Dr. Silvia Restrepo (Advisor) and Dr. Diego Riaño and Dr. Marco Cristancho (Co-advisors) for everything and especially for helping me travel to the to the OSG Summer Course. To my OSG mentor, Tim Cartwright, who was patient enough to answer all the questions from a biologist, To Alain Roy and Igor Sfiligoi, whom taught me a lot and to the OSG staff and Wisconsin University for the grant, the great hospitality and the most beautiful campus I´ve ever known.

References

Beerli, P. (2006). Comparison of Bayesian and maximum-likelihood inference of population genetic parameters. Bioinformatics *22*, 341-345.

Darwin, C. (1859). On the origin of species by means of natural selection (London,, J. Murray).

Drummond, A.J., Nicholls, G.K., Rodrigo, A.G., and Solomon, W. (2002). Estimating Mutation Parameters, Population History and Genealogy Simultaneously From Temporally Spaced Sequence Data. Genetics *161*, 1307-1320.

Guindon, S., Dufayard, J.F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology *59*, 307-321.

Huelsenbeck, J.P., and Ronquist, F. (2005). Bayesian analysis of molecular evolution using MrBayes. Statistical Methods in Molecular Evolution Springer.

Müller, K.F. (2005). The efficiency of different search strategies in estimating parsimony jackknife, bootstrap, and Bremer support. BMC Evol Biol *5*, 58.

Nylander, J.A.A., Ronquist, F., and Huelsenbeck, J.P. (2004). Bayesian Phylogenetic Analysis of Combined Data. Syst Biol.

Omland, K., Cook, L., and Crisp, M. (2008). Tree thinking for all biology: the problem with reading phylogenies as ladders of progress. Bioessays *30*, 854-867.

Stamatakis, A., Hoover, P., and Rougemont., J. (2008). A Rapid Bootstrap Algorithm for the RAxML Web-servers. Systematic Biology *75*, 13.