Let's have a drink now!
HTC Bayesian Phylogenomics on the Zenobe Tier-1 supercomputer

Denis BAURAIN

InBioS / Dept of Life Sciences
University of Liège

CÉCI scientific day / Apr 22, 2016
Background

What is the general problem people in your field of research are trying to solve?
We are trying to reconstruct the Tree of Life.
Objectives

What is the specific problem in that context that you are trying to solve and why does that matter?
We are interested in the evolution of vertebrates, which are the animals having a backbone.
Methods

What strategy are you implementing to solve that problem?
Molecular phylogenetics works by comparing the same gene(s) across different species.
The historical signal lies in the substitutions inherited from the common ancestors of the species.
Multiple substitutions at the same sites erase this signal and even create spurious identities.
We use Markov models of sequence evolution to account for the multiple substitutions.
Phylogenomics is the use of genome-scale data to reconstruct evolutionary relationships.
Animal Sampling

24 freshly sacrificed individuals
RNA Extraction

TRIzol-based
Next-Generation Sequencing

Illumina HiSeq 2000
(or MiSeq)
Dataset Assembly

incl. sequence data from 76 more species
Supermatrix

100 sp. x 2,000,000 sites (4500 genes)
4500 genes (25 shown) / 2,000,000 sites

100 species

zoom on 2 concatenated gene alignments
Methods

Where do computers fit in that strategy? How do you (ab)use them?
Our supermatrix was too large to be analysed "as is". We thus used a gene resampling approach.
Each jackknife replicate was then analysed using a powerful Bayesian phylogenetic software: PhyloBayes MPI.
\[ f(\theta|X) = \frac{f(\theta)f(X|\theta)}{f(X)} \quad \text{with } \theta = (\tau, \nu) \]

\[ f(X) = \int f(\theta)f(X|\theta)d\theta \]

\[ = \sum_{\tau} \int_{\nu} f(\nu)f(X|\tau, \nu)d\nu \]

*Bayes' theorem applied to phylogenetics*

The tree is the interesting part of the model while the model of sequence evolution is a necessary "nuisance".

We want to compute the (posterior) probability distribution of the model (tree and sequence model) given the data (supermatrix).
Models of sequence evolution differ in their capacity to handle subtleties of the evolutionary process.
The best sequence models use complex mixtures that can only be implemented in a Bayesian framework.

The principle is to sample the posterior distribution using numerical simulation (Markov Chain Monte Carlo).
The chain is left running for days until it reaches convergence. Being memoryless, it is naturally restartable at will.
The mean (variance, C.I. etc) of any parameter (including the tree) can then be computed by averaging over samples.
The MCMC cycles over two phases, one to update the topology and another to update the remaining model parameters.
Results

What results have computers enabled you to achieve?

We obtained a new reference phylogenetic framework for the evolution of vertebrates.
We obtained a new reference phylogenetic framework for the evolution vertebrates.
What experience do you feel can be transposed to another research field? Can you share any tips and tricks?
Even if we do not use many cores per job (96), we still need quite a lot of computing power (hence HTC).
Acknowledgment

• Liège
  • Luc Cornet
  • Raphaël Léonard
  • Damien Sirjacobs
• CÉCI
  • David Colignon
• CenAero
  • Danielle Coulon
• Région wallonne
  • Gr. agr. 1117545

• Konstanz
  • Iker Irisarri
• Braunschweig
  • Henner Brinkmann
  • Jörn Petersen
  • Miguel Vences
• Montpellier
  • Frédéric Delsuc
• Moulis / Montreal
  • Hervé Philippe
Let's have a drink now!