Efficient computation of a geodesic distance approximation in phylogenetic tree space

Battagliero S, Puglia G, Vicario S, Rubino F, Scioscia G, Leo P

Motivation

The raising use of phylogenetic studies to address fundamental issues in biology is giving prominence to the need of reliable and efficient tools. Many phylogenetic approaches require tools to compare phylogenetic trees. In fact, several methods have been proposed to build such trees, raising the necessity to compare these results. Moreover, even a single method can give several plausible trees: Bayesian inference methods, for instance, give a sample of trees taken from the posterior distribution, given a sequence dataset. An effective and efficient tree comparison tool would also allow better diagnostic and monitoring of the process of numerical estimation of phylogenetic inference both under a maximum likelihood and a Bayesian framework. We stress the need for efficiency because it is increasingly clear in biology that interesting questions that have a phylogenetic component need to work on large data set, from several hundreds to several thousand individuals, in order to give satisfactory answers. The geodesic tree distance is an effective and conceptually simple way to compare trees, accounting simultaneously for topology and branch lengths. However, even the faster algorithm to compute it, the GTP algorithm, does not scale well to large trees. So we propose a more efficient algorithm, GeoHeuristic, which computes an approximation of the distance, comparing it with GTP and with other approximations, such as the cone path.

Methods

The GeoHeuristic algorithm was tested using two reference phylogenetic tree datasets: the first one contains randomly generated trees and the second one contains trees generated by Bayesian methods. We decided to produce two different datasets, in order to test our algorithm both in a “practical” and in a “difficult” case: the Bayesian tree dataset is a practical case, because it contains trees obtained from real biological data, while the random tree dataset is a difficult case. To analyze the behavior of our algorithm with different numbers of taxa, for both datasets we created 12 groups of trees with increasing number of taxa, from 50 up to 600 taxa with a step of 50. Each group contains 100 pairs of trees for which the distance was computed. We obtained the random tree dataset using Mesquite under a simple uniform probability speciation (Yule) process. We used

[1] IBM Italia S.p.A., GBS BAO Advanced Analytics Services and MBLab, Bari, Italy [2] Consiglio Nazionale delle Ricerche - Istituto di Tecnologie Biomediche, sezione di Bari

the MrBayes software, version 3.2, for phylogenetic inference executing two runs up to 1 million Markov chain generations with a sampling frequency of 10000, obtaining 100 trees per run. The 100 tree pairs were generated matching the two trees with the same generation number from each run. To compare GeoHeuristic algorithm time and memory efficiency with that of GTP we performed time tests on a single CPU of a dual-core Centrino T9300 2.50 GHz with 2GB RAM, while memory allocation tests were performed on a IBM Blade server with 64bit Red Hat Linux operating system. Both algorithms have been implemented in interpreted programming languages (Python and Java, respectively).

Results

According to our experiments, GeoHeuristic relative error differs from that of the cone path from one to three orders of magnitude. In particular, GeoHeuristic attains a relative error always lower than 10-4, corresponding to an accuracy of more than 99.99%. So, GeoHeuristic performed as a very accurate geodesic computation algorithm for most of possible applications. In the case of the cone path, indeed, absolute results are not so bad (relative error between 10-2 and 10-3) for random trees, but are worse for Bayesian trees (always higher than 10-2), which are just the most interesting trees for applicative purposes. Taking the cone path as an example of distance that does not take in consideration split incompatibilities, we learned that split incompatibilities affect significantly, though not dramatically, the geodesic distance. Computation times of GTP and GeoHeuristic for both random trees and Bayesian ones were also measured. In both cases, the GTP time has a much more rapid growth rate than GeoHeuristic time. For the considered window of taxa, the GTP algorithm seems to have a quadratic course, while GeoHeuristic shows an approximate linear one, though theoretical complexities are O(n^4) and O(n^3), respectively. Moreover, the absolute computation time of the GeoHeuristic algorithm is always below 700 milliseconds, even with 600 taxa, for which the GTP algorithm takes more than 7 seconds.

Contact e-mail

simone.battagliero@it.ibm.com