C. R. Palevol 12 (2013) 381–387

Contents lists available at ScienceDirect

Comptes Rendus Palevol

www.sciencedirect.com

General palaeontology, systematics and evolution (Phylogenetic analysis)

Evaluating strategies of phylogenetic analyses by the coherence of their results

Évaluation des stratégies d'analyse phylogénétique par la cohérence de leurs résultats

Blaise Li

Centro de Ciências do Mar, Universidade do Algarve, Campus de Gambelas, 8005-139 Faro, Portugal

ARTICLE INFO

Article history:
Received 30 April 2013
Accepted after revision 4 July 2013
Available online 12 September 2013

Keywords:
Chloroplasts
Coherence
Cyanobacteria
Methods
Phylogeny

Mots clés :
Chloroplastes
Cohérence
Cyanobactéries
Méthodes
Phylogénie

ABSTRACT

I propose an approach to identify, among several strategies of phylogenetic analysis, those producing the most accurate results. This approach is based on the hypothesis that the more a result is reproduced from independent data, the more it reflects the historical signal common to the analysed data. Under this hypothesis, the capacity of an analytical strategy to extract historical signal should correlate positively with the coherence of the obtained results. I apply this approach to a series of analyses on empirical data, basing the coherence measure on the Robinson–Foulds distances between the obtained trees. At first approximation, the analytical strategies most suitable for the data produce the most coherent results. However, risks of false positives and false negatives are identified, which are difficult to rule out.

RÉSUMÉ

Je propose une approche pour identifier, parmi plusieurs stratégies d'analyse phylogénétique, celles aux résultats les plus fiables. Cette approche se base sur l'hypothèse que, plus un résultat est reproduit à partir de données indépendantes, plus il reflète le signal historique commun aux données analysées. Sous cette hypothèse, la capacité d'une stratégie d'analyse à extraire le signal historique devrait être positivement corrélée à la cohérence des résultats obtenus. J'applique cette approche à une série d'analyses sur des données empiriques, en basant la mesure de cohérence sur les distances de Robinson–Foulds entre les arbres obtenus. En première approximation, les stratégies d'analyse les plus adaptées aux données produisent les résultats les plus cohérents. Cependant, des risques de faux positifs et de faux négatifs, difficiles à écarter, sont identifiés.

1. Introduction

An important breakthrough for molecular phylogeny reconstruction has been made with the introduction of probabilistic approaches (Felsenstein, 1981; Yang and Rannala, 1997), directly and explicitly using molecular evolution models. This usually reduces the occurrences of reconstruction artifacts, in particular in studies at large evolutionary scales (but see Simmons, 2012). In parallel with an increased availability of data (which permits a better estimation of the parameters of complex models) and

E-mail address: blaise.li@normalesup.org

1631-0683/$ – see front matter © 2013 Académie des sciences. Publié par Elsevier Masson SAS. Tous droits réservés.
http://dx.doi.org/10.1016/j.crpv.2013.07.001

computational power (which permits the exploration and evaluation of a large number of possible trees), the development of probabilistic methods was accompanied with the development of models that take into account an increasing number of aspects of molecular evolution such as evolutionary rate (Yang, 1993) or composition (Foster, 2004; Lartillot and Philippe, 2004) heterogeneities. The accuracy of phylogenies can also be enhanced by using character selection or recoding techniques (Brinkmann and Philippe, 1999; Goremykin et al., 2010; Hassanin et al., 2005; Inagaki et al., 2004; Roure and Philippe, 2011).

However, the diversity of methods and models available makes it difficult to decide which strategy to adopt when trying to reconstruct a phylogeny. Some methods are available to help the phylogeneticist in this choice. For instance, programs like jModelTest (Posada, 2008) use a variety of criteria to select a model achieving a good compromise between realism and tractability. But such readily available tools are limited to the set of models implemented in the phylogeny programs on which they rely. It is also common practice to compare phylogenies obtained using different models by applying selection criteria identical to those used in *a posteriori* model selection programs, which extends these selection approaches to arbitrary models. Still, the model is only one aspect of the analytical strategy: Data selection or recoding techniques also need to be chosen prior to the tree construction, a program and its specific settings have to be chosen, and support evaluation procedures can take diverse forms. All of these aspects form the analytical strategy that leads from the raw data to an annotated tree ready for drawing phylogenetic conclusions.

An approach suitable for the choice of such integrated analytical strategies could be to make the choice *a posteriori*, based on their results. A variety of analyses would be performed, and the ones producing the most accurate results would be chosen. This immediately raises the question as to how to evaluate the accuracy of a phylogeny reconstruction. Measures such as bootstrap proportions (Felsenstein, 1985) or Bayesian posterior probabilities are sometimes regarded as reliability indicators, but they must be interpreted in the limited context of the particular dataset that has been analysed. Other datasets may yield different support values (or even contradictory results) and these values do not correlate perfectly with one another (Douady et al., 2003). Reliability of phylogenetic relationships is arguably better estimated when considering trees obtained from several independent datasets, and examining the degree to which the results are reproduced across these datasets (Chen et al., 2003; Dettai and Lecointre, 2004; Li and Lecointre, 2009; Miyamoto and Fitch, 1995). In this context, it has been observed that the reproducibility of the results was higher when a better modelling of the data was used (Miyamoto et al., 1994). This justifies a widespread practice consisting in using more complex models and methods when the phylogeny appears more challenging to resolve. This also suggests that result coherence could indeed correlate positively with accuracy.

The purpose of the present article is to report an attempt to use the *a posteriori* approach for selecting strategies of phylogenetic analyses using the reproducibility of the results as a criterion, and to discuss some potential pitfalls of such an approach.

2. Materials and methods

2.1. Test data

The *a posteriori* approach was tested on empirical multigene data assembled in the ambit of a yet-to-be-published work on the phylogeny of Cyanobacteria and plastids (Li et al., in preparation). Given the large evolutionary scale, as well as the potential existence of horizontal gene transfers, such a dataset should provide enough reconstruction challenge so that different analytical strategies will have different reconstruction accuracies, and show various degrees of result coherence.

The data consists of 73 protein-coding genes from 42 Cyanobacteria, plastids or nuclear genes of plastidial origin. The genes were grouped in 4 sets that were considered internally congruent and between them incongruent by the concaterpillar program (Leigh et al., 2008). This program performs a series of likelihood ratio tests under a GTR + I + Γ model, to evaluate whether datasets can be forced to share topologies and branch lengths or if separate trees provide a significantly better likelihood. Results of maximum likelihood analyses under a GTR + I + Γ model should therefore provide a reference situation where some incoherence effectively appears between the datasets. More accurate strategies than maximum likelihood analysis under a GTR + I + Γ model might be able to recover more of the history common to all datasets, for each one of them, and therefore be characterised by a higher coherence in the results.

2.2. Analytical strategies tested

For each of the 4 combined datasets, a series of various analytical strategies were applied. A name is associated with each of them to facilitate reporting and discussion of the results.

Maximum likelihood bootstrap analyses were conducted using RAxML versions 7.0.4 and 7.3.4 (Stamatakis, 2006) under a GTR + I + Γ model, with 200 pseudo-replicates of the data. For these analyses, the original data matrices were used, their amino-acid translations (for which a CPREV + I + Γ model was used) as well as some versions of these matrices where diverse combinations of sites were subjected to codon-degeneracy recodings.

A codon-degeneracy recoding is based on the replacement of codons by degenerate versions that represent all codons coding the same amino-acid. Nucleotides are replaced by IUPAC ambiguity codes at codon positions where several codons for the same amino-acid differ.

The goal of these recodings is to eliminate potentially misleading signal. The signal considered for removal corresponds to sites involved in codon synonymy. Due to the relaxed selection on the nucleotide at such sites, convergence between sequences sharing the same bias in their genome's nucleotide composition may have happened and mislead phylogenetic reconstruction (see for instance Cox et al., 2008; Foster, 2004; Hassanin et al., 2005; Nabholz et al., 2011; Rota-Stabelli et al., 2013). The most useful of