Available online at www.sciencedirect.com

SCIENCE DIRECT

Molecular Phylogenetics and Evolution 40 (2006) 428–434

MOLECULAR PHYLOGENETICS AND EVOLUTION

www.elsevier.com/locate/ympev

Phylogenetic estimation under codon models can be biased by codon usage heterogeneity

Yuji Inagaki [a,*], Andrew J. Roger [b]

[a] *Center for Computational Sciences, Institute of Biological Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8577, Japan*
[b] *Genome Atlantic and Canadian Institute for Advanced Research, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia, Canada B3H 1X5*

Received 6 October 2005; revised 10 March 2006; accepted 15 March 2006
Available online 2 May 2006

Abstract

In theory, codon models that account for the dependence of nucleotide substitutions between codon positions as well as differences between synonymous and non-synonymous changes best describe the sequence evolution in protein coding genes. However, in practice we know little about the degree to which violations of the assumptions of codon model-based estimates occur, and how significant these artifacts may be. In nucleotide-based phylogenies from first and second codon positions in a concatenated plastid gene data set, two distantly related taxa—dinoflagellate and haptophyte plastids—were robustly grouped together. This artifactual grouping is attributed to the parallel heterogeneity in leucine (Leu) and serine (Ser) codon usages in the data set. Here, by using this data set, we demonstrated that codon-based phylogenetic estimations are seriously biased, robustly uniting the dinoflagellate and haptophyte plastids into a monophyletic clade, when the model assumption of homogeneity of codon composition was violated. Our results suggest that similar phylogenetic artifacts may occur via codon usage heterogeneity in any amino acids in codon model-based estimations. We advise that homogeneity in codon usage across taxa in a data set be confirmed before codon model-based phylogenetic estimation is attempted.

Keywords: Model misspecification; Long-branch attraction; Phylogenetic artifact; Chromist; Dinoflagellates; Plastid evolution

1. Introduction

Nucleotide evolution in protein-coding regions is dependent on codon structure, since selection on nucleotide substitutions operates at the codon (or amino acid [aa] sequence) level rather than at the single-nucleotide level. Due to the degeneracy of the genetic code, codon positions where synonymous changes are possible rapidly accumulate multiple substitutions. If the genomic region including a particular gene possesses a unique base/codon composition, such characteristics may heavily affect synonymous substitutions (Duret, 2002; Graur and Li, 2000). This "non-stationary" process of nucleotide change can produce heterogeneous base/codon composition across a phylogenetic tree. Nucleotide-based models assuming the homogeneity of base/codon composition [e.g., general-time-reversible (GTR)[1] models] may not sufficiently describe the nucleotide evolution in protein-coding genes, and phylogenetic estimates under these models could be significantly biased (Foster and Hickey, 1999; Galtier and Gouy, 1998; Yang and Roberts, 1995). To counter this problem, the

[*] Corresponding author. Fax: +81 29 853 6406.
E-mail address: yuji@ccs.tsukuba.ac.jp (Y. Inagaki).

[1] *Abbreviations used:* GTR, general-time-reversible; R, purine; Y, pyrimidine; DNA^{1+2} analysis, Nucleotide-based analysis of first and second codon positions; BP, bootstrap; H + D plastid clade, haptophyte, and dinoflagellate plastid clade; S + D plastid clade, stramenopile, and dinoflagellate plastid clade; PsaA, photosystem I P700 chlorophyll *a* apoprotein A1; PsbA, photosystem II D1 protein; constant L/S (or L/S/R) codon sites, the codon sites fixed to Leu or Ser (or Leu, Ser or Arg) in the corresponding amino acid data set.

1055-7903/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.ympev.2006.03.020

LogDet/Paralinear (LogDet) distance method was established and is widely used for nucleotide-based phylogenies. Alternatively, base composition heterogeneity in nucleotide data sets can be lowered by re-coding pyrimidines (T and C) and purines (A and G) into Y and R, respectively ["RY-coding method" (Harrison et al., 2004; Phillips and Penny, 2003; Phillips et al., 2004, 2001)].

Codon models can account for both dependence of nucleotide changes within a codon and synonymous versus non-synonymous substitutions (Goldman and Yang, 1994), so these models should be more appropriate for phylogenetic analyses of protein-coding genes than nucleotide-based models. Codon models have been implemented in major phylogenetic software packages [e.g., PAML (Yang, 1997) and MRBAYES (Ronquist and Huelsenbeck, 2003)], and have recently been recommended as the most appropriate models for phylogenetic analyses of coding regions (Ren et al., 2005; Shapiro et al., 2006). Importantly, MRBAYES now allows estimation of optimal trees from codon data with codon models using efficient Monte Carlo Markov Chain (MCMC) sampling procedures. One major difficulty in codon model-based analyses is extensive cost in computation. Maximum likelihood (ML) or Bayesian analyses utilizing codon models, perform calculations using matrices of 3721 (61 × 61) transition probabilities, and therefore are much more computationally expensive than analyses using nucleotide or aa models with much smaller matrices (4 × 4 or 20 × 20 matrices, respectively). Fortunately, current desktop computers are sufficiently fast to run codon analyses on small to medium size data—e.g., in this study, the most intensive Bayesian codon analyses of a 19-taxon data set took ~160 h to complete by single Mac G5 2.0 GHz processor (see below for details). Although the popularity of codon models is growing, the conditions under which analyses employing codon models yield biased estimates remain to be investigated. Of particular interest is the impact of codon usage heterogeneity amongst the sequences under consideration, since, to our knowledge, all codon models currently available assume the homogeneity of codon composition.

In plastid *psbA* genes (encoding photosystem II D1 protein), some peridinin-type dinoflagellate and haptophyte plastids appear to share the unique and idiosyncratic arginine (Arg), leucine (Leu), and serine (Ser) codon usage patterns (Henceforth peridinin-type dinoflagellate plastids will be referred simply as "dinoflagellate plastids"). We have recently shown that, as no nucleotide-based models can account for such parallel codon usage heterogeneity, the nucleotide-based phylogenetic estimates from the *psbA* data set are seriously biased, recovering a highly supported, but artifactual clade of haptophyte and dinoflagellate plastids (H + D plastid clade) (Inagaki et al., 2004). In contrast, analyses of the corresponding protein data set does not support the H + D plastid clade (Inagaki et al., 2004). Recent phylogenetic analyses of photosystem II component CP43 (PsbC) and nuclear-encoded plastid-targeted glyceraldehydes-3-phosphate dehydrogenase also disfavor a grouping of haptophyte and dinoflagellate plastids (Takishita et al., 2004, 2005). A monophyletic group of the stramenopile and dinoflagellate plastids (S + D plastid clade) was recovered by phylogenetic analyses of concatenated protein data sets comprised of *psaA* plus *psbA* genes, and *psaA* plus *psaB* plus *psbA* plus *psbC* plus *psbD* genes (Inagaki et al., 2004; Yoon et al., 2005). Importantly, the S + D plastid clade recovered by these plastid phylogenies are consistent with other phylogenetic analyses of nuclear-encoded genes (Cavalier-Smith, 1999; Fast et al., 2001; Harper and Keeling, 2003; Harper et al., 2004; Simpson et al., 2006; Takishita et al., 2005). Recent multi-gene phylogenetic analyses of 10 plastid genes (including *psbA*) reconstructed the H + D plastid clade with relatively high statistical support, but the bootstrap values were largely dependent on the presence versus absence of the *psbA* gene (Bachvaroff et al., 2005). These results confirmed that the H + D plastid clade in *psbA* phylogenies is likely a data set-specific artifact.

The plastid gene data sets that include the dinoflagellate and haptophyte plastid sequences provide an ideal system to evaluate the robustness of codon-based phylogenetic analyses to codon usage heterogeneity. In this study, we have confirmed that the artifactual allegiance of the dinoflagellate and haptophyte plastids can be attributed to the parallel codon bias shared between the two plastids in a concatenated *psaA* + *psbA* data set (the former gene encodes photosystem I P700 chlorophyll *a* apoprotein A1). Then, by using this concatenated data set, we illustrate that codon-model based phylogenetic estimation can be positively misleading when the assumption of the homogeneity of codon composition across sequences is violated.

2. Materials and methods

2.1. Amino acid-based phylogenetic analyses

PsaA and PsbA aa sequences were sampled from 19 plastids of *Cyanophora paradoxa* (a glaucophyte), seven red algae, eight chromists (four haptophyte, two stramenopile, two cryptophytes), and three dinoflagellates. The details of the sequences considered here are described in Supplementary table. By omitting ambiguously aligned sites and sites including gaps, 575 aa sites were subjected to phylogenetic analyses described below ("PsaA + PsbA" data set).

ML trees were estimated under JTT models accounting for among-site rate variation (ASRV) with sequence addition randomized for five times, followed by global rearrangements using PROML implemented in PHYLIP version 3.6a (Felsenstein, 1993). ASRV was modeled by using discrete gamma (Γ) distribution with four equally probable categories. The model parameters were estimated from the data using TREE-PUZZLE version 5.2 (Schmidt et al., 2002). ML bootstrap analyses (100 replicates) were performed under the same settings described above, except with a single randomized sequence addition per replicate.