ABSTRACT

Background: Here, we develop a new codon-based model, which consists of mutations at the nucleotide level and selection at the amino acid level via a genetic code, and estimate selective constraints on amino acids from available empirical substitution matrices. The mutational process of individual codons is modeled as a reversible Markov process, and multiple nucleotide changes are assumed to occur with the same order of time as single nucleotide changes do. In a codon substitution model, 1829 codon exchangeabilities must be determined. In the present model, they are expressed as functions of 6 nucleotide mutation rates and selective constraints for 190 types of amino acid replacements. If the selective constraints are estimated, and their relative strengths among amino acid replacements are approximated to be constant irrespective of protein families, parameters to be optimized in maximum likelihood (ML) and Bayesian inferences of phylogenetic trees will be drastically reduced to parameters for nucleotide mutations and some additional ones.

Results: The present model with substitution rates that are assumed to obey a Γ distribution can be well fitted to each 1-PAM matrix of empirical amino acid substitution matrices (JTT, WAG, and LG) and empirical codon substitution matrix (KHG) already published. ML estimators of selective constraints on amino acids are calculated together with other parameters. Akaike information criterion (AIC) values indicate that the assumption of multiple nucleotide changes significantly better fits the model to the empirical substitution matrices. One of interesting results is that the ML estimators of transition to transversion bias obtained from these empirical matrices are not so large as previously estimated. Also, the present model with the selective constraints estimated from the JTT/WAG/LG/KHG can be well fitted to other matrices including the ones (cpREV) for chloroplast proteins and (mtREV) for vertebrate mitochondrial proteins.

Conclusions: Thus, the present codon-based model with the ML estimators for the selective constraints and with adjustable mutation rates of nulceotides would be useful as a simple substitution model in ML and Bayesian inferences of molecular phylogenetic trees, and enables us to to obtain biologically meaningful information at both nucleotide and amino acid levels from codon and protein sequences.