Advantages of a Mechanistic Codon Substitution Model for Evolutionary Analysis of Protein-Coding Sequences

Sanzo Miyazawa*

Graduate School of Engineering, Gunma University, Kiryu, Gunma, Japan

Abstract

Background: A mechanistic codon substitution model, in which each codon substitution rate is proportional to the product of a codon mutation rate and the average fixation probability depending on the type of amino acid replacement, has advantages over nucleotide, amino acid, and empirical codon substitution models in evolutionary analysis of protein-coding sequences. It can approximate a wide range of codon substitution processes. If no selection pressure on amino acids is taken into account, it will become equivalent to a nucleotide substitution model. If mutation rates are assumed not to depend on the codon type, then it will become essentially equivalent to an amino acid substitution model. Mutation at the nucleotide level and selection at the amino acid level can be separately evaluated.

Results: The present scheme for single nucleotide mutations is equivalent to the general time-reversible model, but multiple nucleotide changes in infinitesimal time are allowed. Selective constraints on the respective types of amino acid replacements are tailored to each gene in a linear function of a given estimate of selective constraints. Their good estimates are those calculated by maximizing the respective likelihoods of empirical amino acid or codon substitution frequency matrices. Akaike and Bayesian information criteria indicate that the present model performs far better than the other substitution models for all five phylogenetic trees of highly-divergent to highly-homologous sequences of chloroplast, mitochondrial, and nuclear genes. It is also shown that multiple nucleotide changes in infinitesimal time are significant in long branches, although they may be caused by compensatory substitutions or other mechanisms. The variation of selective constraint over sites fits the datasets significantly better than variable mutation rates, except for 10 slow-evolving nuclear genes of 10 mammals. An critical finding for phylogenetic analysis is that assuming variable mutation rates over sites lead to the overestimation of branch lengths.

Citation: Miyazawa S (2011) Advantages of a Mechanistic Codon Substitution Model for Evolutionary Analysis of Protein-Coding Sequences. PLoS ONE 6(12): e28892. doi:10.1371/journal.pone.0028892

Editor: Darren P. Martin, Institute of Infectious Disease and Molecular Medicine, South Africa

Received September 14, 2011; **Accepted** November 16, 2011; **Published** December 29, 2011

Funding: The authors have no funding or support to report.

Competing Interests: The author has declared that no competing interests exist.

* E-mail: sanzo.miyazawa@gmail.com

Introduction

Growing DNA and protein sequence data is now a valuable source of knowledge in many fields of science, especially in evolutionary biology. Evolutionary history of DNA sequences is a key to understand the diversity of homologous sequences. Any method for inferring molecular phylogeny is implicitly or explicitly based on the evolutionary model of nucleotide or amino acid substitutions, and the reliability of phylogenetic analyses strongly depends on models designed to approximate the substitution processes of nucleotide and amino acid. For the evolutionary analysis of protein-coding sequences, three types of models can be used: nucleotide, amino acid, and codon substitution models. Which type of model fits any sequence data better than the others? Mutational events occur at the nucleotide level, but selective pressure primarily operates at the amino acid level. Thus, a codon substitution model has a potential to outperform both nucleotide substitution models [1–3] and amino acid substitution models [4–12], because it can take into account both mutational tendencies at the nucleotide level and selective pressure on amino acid replacements as well as a genetic code. Shapiro et al. [13] proposed a codon position model, in which codon position is incorporated into a nucleotide substitution model. This model is computationally efficient but insufficient to take account of the dependencies of selective pressure on amino acid replacements.

Codon substitution models are classified into either an empirical codon substitution model or a mechanistic codon substitution model. In empirical codon substitution models [14,15], substitution rates between codons were empirically estimated from a large set of protein-coding sequences, and mutational tendencies at the nucleotide level and selection pressure at the amino acid level cannot be separated at all. Therefore, there is no parameter except codon frequencies to tailor for each protein family. Delport et al. [16] showed that empirical substitution matrices represent the average tendencies of substitutions over various protein families by sacrificing gene-level resolution.

In mechanistic codon substitution models, a mutational mechanism at the nucleotide level and selection at the amino acid level are distinguished in various levels of separation. If no selection pressure on amino acids is taken into account, the codon substitution model will become essentially equivalent to a nucleotide substitution model. If mutation rates are assumed not to depend on the codon type, then the model will become essentially equivalent to an amino acid substitution model. Such a codon model with the infinitely large synonymous substitution rate, was proved [17] to be exactly equivalent to an amino acid substitution model. It was shown on protein-coding sequences that codon substitution models are statistically superior to the nucleotide and amino acid substitution models [18,19].

There are two type of models for the mutational scheme of codon, depending on whether multiple nucleotide changes in infinitesimal time are allowed [17,19–21] or not. Even though all the empirical amino acid substitution models [4–6,8,10,11] and the empirical codon substitution model [15] allow amino acid or codon substitutions requiring multiple nucleotide changes in infinitesimal time, only single nucleotide changes were assumed to occur in infinitesimal time [7,18,22–25]. Multiple nucleotide changes in infinitesimal time are biologically plausible, because they can be caused by successive compensatory substitutions [26], recombination, gene conversion and other mechanisms [27], especially in long branches. It has been pointed out that assuming multiple nucleotide changes in codon substitution models can significantly improve the maximum likelihood (ML) value [19,20].

In the present models, mutational tendencies at the nucleotide level are tailored to each gene by the general time-reversible (GTR) model, but multiple nucleotide changes in infinitesimal time are allowed. In the Singlet-Doublet-Triplet (SDT) mutation model [20], single-nucleotide, doublet and triplet mutations spanning codon boundaries are taken into account, but double nucleotide mutations at the first and the third positions in a codon were not taken into account. In the present model, it is assumed [19] that nucleotide mutations occur independently at each codon position and so any double nucleotide mutation occurs as frequently as doublet mutations.

There are a variety of models for selection pressure on amino acid replacements in mechanistic codon substitution models; (1) models [17,18,21] based on empirical amino acid substitution matrices, in which codon exchangeabilities for nonsynonymous substitutions were evaluated on the basis of empirical amino acid exchangeabilities, and selective constraints on amino acids are not well separated from codon mutation rates, (2) equal-constraint models [20,24,28,29], in which the difference between nonsynonymous and synonymous substitution rates was taken into account but the amino acid dependences of selective constraints on amino acids were not taken into account, i.e., single selective constraints for all types of amino acid substitutions, (3) physico-chemical-constraint models [7,22,23], in which selective constraints for each protein family were approximated in a linear function of the selective constraints evaluated from physico-chemical properties of amino acids, (4) fully-parameterized-constraint models [7,16,25], in which selective constraints were grouped, and the number of groups and the strength of selective constraint of each group were optimized for a given protein phylogeny, and (5) site-specific selection models [30], in which site-specific selection was modeled in terms of site-specific residue frequencies in a codon substitution model.

In the models [17,21] of the first category, codon exchangeabilities for nonsynonymous codon substitutions requiring multiple nucleotide changes are set to non-zero according to the empirical amino acid exchangeabilities; the exchangeability is defined to be an instantaneous rate divided by the equilibrium composition of destination codon or amino acid. The method in the fourth category has the highest resolution of selective constraints employing as many substitution groups as necessary. However, it seems to be a very computer-intensive calculation [25].

In the present model, selective constraints on the respective types of amino acid replacements are tailored to each gene in a linear function of a given estimate of selective constraints in the same way with the physico-chemical-constraint models. The simplest model for the selective constraints is to assume equal constraint on amino acid replacements and equivalent to the second category of model; it is named here the Equal-Constraint model. Of course, physico-chemical estimates of the selective constraints can also be used [19]. Better estimates are those that were estimated [19] by maximizing the respective likelihoods of observed amino acid or codon substitution frequency matrices.

A property of codon substitution models in which synonymous substitutions can be identified is an advantage over nucleotide and amino acid substitution models. Significance of rate variation over sites in proteins has been demonstrated mostly in nucleotide substitution models and empirical amino acid substitution models [31–33]. Variable rates of nucleotide and amino acid substitutions over sites can be caused not only by the variation of mutation rate but also by the variation of selective constraint over sites. However, in the nucleotide and the amino acid substitution models, synonymous substitutions cannot be recognized, and therefore the variations of mutation rate and of selective constraint over sites cannot be distinguished from each other. On the other hand, the variations of selective constraint and mutation rate can be distinguished from each other in codon substitution models, assuming no selective pressure on synonymous mutations at the amino acid level. It is reasonable from a viewpoint of protein structure and function that amino acid replaceabilities strongly depend on sites in a protein [34]. Molecular mechanisms are not known to cause significantly variable mutation rates over sites within the exons of a gene. Here, we examine which model fits data statistically better. In the present model, either the variation of mutation rate or the variation of selective constraint is taken into account, although both is not taken into account at the same time because of heavy computational load. Yang et al. [29] also studied heterogeneous selection pressure at amino acid sites by codon substitution models.

Besides the variation of substitution rate over sites, the variation of substitution rate over time at each site is also possible. The site-specific variation of substitution rate over time was first discussed as a covarion model by Fitch and Markowitz [35]. Recently, a few cases indicating its significance have been reported [36,37]. Here we take into account the variation of mutation rate over time at each site in a simple approximation.

The estimation of branch lengths is critical on the estimation of phylogeny and divergence times. We examine how differently branch lengths are estimated between models. The present mechanistic codon substitution model can simulate a wide range of codon substitution processes by changing parameters, and can provide biologically meaningful information at both nucleotide and amino acid levels such as transition/transversion rate bias, the ratio of multiple nucleotide changes, the strength of selective constraints on amino acids, the variation of mutation rate or selective constraints over sites, and also the variation of mutation rate over time in branches. Here, the present codon substitution models with the various sets of parameters are extensively studied, and the advantages of the present model over other models are demonstrated.

Methods

A time-reversible Markov model for substitutions

When substitutions independently occur at each site with a constant substitution rate $R_{\kappa\lambda}$ per unit time from codon or amino acid κ to λ, the substitution probability matrix $S(t)$ at time t is calculated as

$$S(t) = \exp(Rt) \qquad (1)$$

Assuming that the detailed balance condition between states is satisfied, i.e., $f_\kappa S(t)_{\kappa\lambda} = f_\lambda S(t)_{\lambda\kappa}$ and $f_\kappa R_{\kappa\lambda} = f_\lambda R_{\lambda\kappa}$, the substitution rate $R_{\kappa\lambda}$ is represented as

$$R_{\kappa\lambda} = r_{\kappa\lambda} f_\lambda \quad , \quad r_{\kappa\lambda} = r_{\lambda\kappa} \quad for\ \kappa \neq \lambda \qquad (2)$$

where f_κ is the equilibrium composition; $\sum_\kappa f_\kappa S(t)_{\kappa\lambda} = f_\lambda$. The symmetric matrix r is named an exchangeability matrix. In the case of the codon substitution matrix, the equilibrium frequencies of stop codons are set to be equal to 0, and therefore the probability flow from any to a stop codon and its inverse flow are always equal to 0. The unit of time is chosen in such a way that the total rate of R is equal to 1;

$$\sum_\kappa f_\kappa \sum_{\lambda \neq \kappa} R_{\kappa\lambda} = -\sum_\kappa f_\kappa R_{\kappa\kappa} = 1 \qquad (3)$$

Therefore, only the relative values among $r_{\kappa\lambda}$ are meaningful.

In a given phylogeny of molecular sequences, a substitution process of codon or amino acid is assumed to be in equilibrium. In other words, the substitution process is assumed to be time-reversible. Also, exchangeabilities $\{r_{\kappa\lambda}\}$ are approximated not to depend on the equilibrium frequencies $\{f_\kappa\}$; this model is specified here with a suffix "F" according to a common naming convention.

Empirical amino acid substitution models converted into codon substitution models

Amino acid exchangeabilities $\{r_{ab}\}$ for amino acid substitutions have been estimated from large sets of protein sequences. From nuclear proteins, the JTT [5], the WAG [10], and the LG [11] rate matrices were estimated. The mtREV [6] substitution probability matrix was estimated from vertebrate mitochondrial proteins, and the cpREV10 [8] and the cpREV64 [38] matrices were estimated from chloroplast proteins of 10 species and of 64 species, respectively.

These amino acid substitution models can be converted into codon substitution models by defining codon exchangeabilities on the basis of amino acid exchangeabilities between encoded amino acids [17–19,21]. Here we consider the following simplest conversion from the amino acid models into codon models to examine the performance of the empirical amino acid substitution models in phylogenetic inference from coding sequences.

The codon exchangeability $r_{\mu\nu}$ between nonsynonymous codons μ and ν is defined to be proportional to the empirical amino acid exchangeability $r_{a_\mu b_\nu}^{empirical}$ between encoded amino acids a_μ and b_ν with a parameter w_0 to adjust the ratio of synonymous to non-synonymous substitution exchangeability. Codon exchangeabilities between synonymous codons are taken to be all equal to one another in such a way that in the case of $w_0 = 0$ they are equal to the maximum exchangeability of nonsynonymous substitutions.

$$r_{\mu\nu} = \begin{cases} C_{\text{onst}}\, r_{a_\mu b_\nu}^{\text{empirical}} e^{w_0} & \text{for } a_\mu \neq b_\nu \\ C_{\text{onst}} \max_a \max_{b \neq a} r_{ab}^{\text{empirical}} & \text{for } a_\mu = b_\nu \end{cases} \qquad (4)$$

where "empirical"∈{JTT, WAG, LG, mtREV, cpREV10, cp REV64}. The arbitrary scaling constant C_{onst} is determined by Eq. 3. In the limit of $w_0 \rightarrow -\infty$, this model is exactly equivalent to the corresponding amino acid substitution model [17]. This model was named as the SK-P1 model by Seo and Kishino [17], and is called here by the name of the empirical amino acid substitution matrix with a suffix meaning the number of ML parameters such as JTT-*n*, WAG-*n*, LG-*n*, mtREV-*n*, cpREV10-*n*, and cpREV64-*n*.

Empirical codon substitution models

Kosiol et al. [15] estimated codon exchangeabilities $\{r_{\mu\nu}\}$ from nuclear-encoded sequences; this substitution rate matrix is called here the KHG matrix. This empirical codon substitution model has been extended here with a parameter w_0 to adjust the ratio of synonymous to non-synonymous substitution exchangeability.

$$r_{\mu\nu} = C_{\text{onst}}\, r_{\mu\nu}^{\text{empirical}} [\delta_{a_\mu b_\nu} + \frac{\max_{\mu\nu}(r_{\mu\nu}^{\text{empirical}} \delta_{a_\mu b_\nu})}{\max_{\mu\nu}(r_{\mu\nu}^{\text{empirical}}(1-\delta_{a_\mu b_\nu}))} \qquad (5)$$
$$(1-\delta_{a_\mu b_\nu})] \exp(w_0(1-\delta_{a_\mu b_\nu}))$$

for $\mu \neq \nu$, where empirical∈{KHG}, and $\delta_{a_\mu b_\nu}$ is the Kronecker's δ. The arbitrary scaling constant C_{onst} is determined by Eq. 3. The exchangeabilities of nonsynonymous codon substitutions are scaled in such a way that in the case of $w_0 = 0$ the maximum exchangeability of nonsynonymous substitutions is equal to that of synonymous substitutions. This model is called KHG-*n*, where the suffix *n* means the number of ML parameters.

A mechanistic codon substitution model with multiple nucleotide changes

In the present mechanistic codon substitution model [19], the substitution rate $R_{\mu\nu}$ is represented as the product of a mutation rate $M_{\mu\nu}$ and the average rate of fixation $F_{\mu\nu}$, which is defined to be the average fixation probability multiplied by the chromosomal population size, for mutations from codon μ to ν under selection pressure; $R_{\mu\nu} \propto M_{\mu\nu} F_{\mu\nu}\ for\ \mu \neq \nu$. The M is also assumed to satisfy the detailed balance condition; $f_\mu^{\text{mut}} M_{\mu\nu} = f_\nu^{\text{mut}} M_{\nu\mu}$, where f_ν^{mut} is the equilibrium codon composition of the rate matrix M. Under this assumption, the average fixation rate $F_{\mu\nu}$ must be represented as the product of the two terms, $f_\nu / f_\nu^{\text{mut}}$ and $e^{w_{\mu\nu}}$, where $w_{\mu\nu} = w_{\nu\mu}$; $F_{\mu\nu} = (f_\nu / f_\nu^{\text{mut}}) e^{w_{\mu\nu}}\ for\ \mu \neq \nu$. Then, the exchangeability $r_{\mu\nu}$ can be represented as

$$R_{\mu\nu} = r_{\mu\nu} f_\nu = C_{\text{onst}}\ M_{\mu\nu} \frac{f_\nu}{f_\nu^{\text{mut}}} e^{w_{\mu\nu}} \quad for\ \mu \neq \nu \qquad (6)$$

The arbitrary scaling constant C_{onst} is is determined by Eq. 3.

The frequency-dependent term $f_\nu / f_\nu^{\text{mut}}$ represents the effects of selection pressures at the DNA level as well as at the amino acid level, which change the codon frequency from the mutational equilibrium frequency f_ν^{mut} to the frequency f_ν specific to a gene. The fixation rate F was explicitly given as a function of the fitnesses of mutants μ and ν [28,30]. The fixation rate is obviously equal to 0 for lethal mutations and equal to 1 for neutral mutations. Here, we approximate the average quantity $e^{w_{\mu\nu}}$ over mutants to be independent of codon frequencies. This quantity $e^{w_{\mu\nu}}$ is essentially the same as the one called the rate of acceptance by Miyata et al. [39]. We assume that selection pressure against codon replacements appears primarily on an amino acid sequence encoded by a nucleotide sequence; in other words, $w_{\mu\nu}$ for codon pair (μ,ν) is equal to the selective constraint $w_{a_\mu b_\nu}$ for the encoded amino acid pair (a_μ, b_ν).

$$e^{w_{\mu\nu}} \equiv \begin{cases} e^{w_{a_\mu b_\nu}} & \text{for} \quad \mu,\nu \notin \{\text{ stop codons }\} \text{ and } \mu \neq \nu \\ 0 & \text{for} \quad \mu\ or\ \nu \in \{\text{ stop codons }\} \text{ and } \mu \neq \nu \end{cases} \qquad (7)$$

At the amino acid level, there should be no selection pressure against synonymous mutations. Thus, the w_{ab} satisfies

$$w_{ab} = w_{ba} \quad , \quad w_{aa} = 0 \tag{8}$$

Selective constraints $w_{\mu\nu}$ are evaluated for a specific protein family in a linear function of a given estimate of w_{ab};

$$w_{ab} \equiv \min[\beta w_{ab}^{\text{estimate}} + w_0(1-\delta_{ab}),0] \tag{9}$$

where w_{ab}^{estimate} with $''\text{estimate}'' \in \{\text{Equal}-\text{Constraint, EI, JTT}-\text{ML91}+, \text{WAG}-\text{ML91}+, \text{LG}-\text{ML91}+, \text{KHG}-\text{ML 200}\}$ means the estimate of w_{ab}, which is equal constraint on amino acids ($w_{ab}^{\text{estimate}} = 0$ or $\beta = 0$), a physico-chemical estimate based on the Energy-Increment-based (EI) method [19], or a ML estimate [19] from the empirical substitution frequency matrix of JTT, WAG, LG, or KHG. The value of w_{ab} is non-positive, assuming that on average there is negative selection on amino acid replacements; of course, $w_{ab}^{\text{estimate}} \leq 0$ [19]. The parameter β, which is non-negative, adjusts the strength of selective constraints for a given protein family. The parameter w_0 directly controls the ratio of nonsynonymous to synonymous substitution exchangeability. Positive selection is taken into account when selective constraints are variable over sites.

The Equal-Constraint model with $w_0 = 0$ is called the No-Constraint model and is equivalent to a nucleotide substitution model. In the model EI, $\hat{w}_{ab}^{\text{estimate}} \equiv \Delta\hat{\varepsilon}_{ab}^{c} + \Delta\hat{\varepsilon}_{ab}^{v}$, where $\Delta\hat{\varepsilon}_{ab}^{c}$ and $\Delta\hat{\varepsilon}_{ab}^{v}$ represent the mean increment of contact energies between residues and the mean volume change due to an amino acid replacement, respectively; see Supporting Information, Text S1, in [19]. The selective constraint matrices $w^{\text{ estimate}}$ with $''\text{estimate}'' \in \{\text{JTT}-\text{ML91}+, \text{WAG}-\text{ML91}+, \text{LG}-\text{ML91}+\}$ were those estimated by maximizing the respective likelihoods of the 1-PAM amino acid substitution frequency matrices of JTT, WAG, and LG in the ML-91+ model [19]. Similarly, the matrix $w^{\text{ KHG}-\text{ML200}}$ were estimated from the 1-PAM KHG codon substitution frequency matrix in the ML-200 model [19]. These estimates of selective constraints are available as Supporting Information, Data S1, in [19]. These models are called here by the name of a selective constraint matrix with a suffix meaning the number of ML parameters such as Equal-Constraint-*n*, EI-*n*, JTT/WAG/LG-ML91+-*n*, and KHG-ML200-*n*.

The mutation rate matrix M is defined in terms of nucleotide mutation rates as follows.

$$M_{\mu\nu} \equiv \prod_{i=1}^{3}[\delta_{\mu_i\nu_i} + (1-\delta_{\mu_i\nu_i})B_{i,\mu_i\nu_i}] \text{ for } \mu \neq \nu \tag{10}$$

where B_i is a mutation rate matrix between the four types of nucleotides at the ith codon position, $\delta_{\mu_i\nu_i}$ is the Kronecker's δ, and the index μ_i means the ith nucleotide in the codon μ; $\mu = (\mu_1,\mu_2,\mu_3)$ where $\mu_i \in \{$ a, t, c, g $\}$. Assuming that the rate matrix B_i satisfies the detailed balance condition, it is represented as

$$B_{i,\mu_i\nu_i} = m_{i,\mu_i\nu_i} f_{i,\nu_i}^{\text{mut}} \qquad \text{for } i = 1,2,3 \tag{11}$$

$$m_{i,\mu_i\nu_i} = m_{i,\nu_i\mu_i} \tag{12}$$

$$f_{\nu=(\nu_1,\nu_2,\nu_3)}^{\text{mut}} = f_{1,\nu_1}^{\text{mut}} f_{2,\nu_2}^{\text{mut}} f_{3,\nu_3}^{\text{mut}} \tag{13}$$

where f_{i,ν_i}^{mut} is the equilibrium composition of nucleotide ν_i at the ith codon position, and $m_{i,\mu_i\nu_i}$ is the exchangeability between nucleotides μ_i and ν_i at the ith codon position. Because the B_i is assumed to satisfy the detailed balance condition, the M also satisfies the detailed balance condition.

If multiple nucleotide changes were completely ignored, then Eq. 10 would be simplified as $M_{\mu\nu} = ((1-\delta_{\mu_1\nu_1})B_{1,\mu_1\nu_1}\delta_{\mu_2\nu_2}\delta_{\mu_3\nu_3}) + (\delta_{\mu_1\nu_1}(-\delta_{\mu_2\nu_2})B_{2,\mu_2\nu_2}\delta_{\mu_3\nu_3}) + (\delta_{\mu_1\nu_1}\delta_{\mu_2\nu_2}(1-\delta_{\mu_3\nu_3})B_{3,\mu_3\nu_3})$, whose formulation for a codon mutation rate matrix with Eq. 11 is the same as the one proposed by Muse and Gault [24]. Here, it should be noted that $B_{i,\mu_i\nu_i}$ in Eq. 11 is defined to be proportional to the equilibrium nucleotide composition f_{i,ν_i}^{mut}. Alternatively, one may define $M_{\mu\nu}$ as $M_{\mu\nu} = \Pi_{i=1}^{3}[\delta_{\mu_i\nu_i} + (1-\delta_{\mu_i\nu_i})m_{i,\mu_i\nu_i}]f_{\nu}^{\text{ mut}}$ in the same way as Miyazawa and Jernigan [22] and others [7,23] defined it to be proportional explicitly to the composition of the base triplet, $f_{\nu}^{\text{ mut}}$. This alternative definition with Eqs. 6 is equivalent to Eqs. 10 and 11 with $f_{\nu_i}^{\text{ mut}} = 0.25$ and $m_{i,\mu_i\nu_i} \Rightarrow 4m_{i,\mu_i\nu_i}$, and thus it is a special case in the present formulation.

The No-Constraint model, in which there is no selection pressure on amino acid replacements ($w_{\mu\nu} = 0$), is a nucleotide substitution model extended to allow multiple nucleotide changes in infinitesimal time. Also, it is useful to note that the present model in the special case of $M_{\mu\nu} = \mathit{constant}$ becomes equivalent to an amino acid substitution model converted into a codon substitution model; if $(m_i)_{\mu_i\nu_i} = 4$ and $f_{i,\nu_i}^{\text{ mut}} = 0.25$, then $M_{\mu\nu} = 1$ and Eq. 6 will become $r_{\mu\nu} \propto e^{w_{\mu\nu}}$ and equivalent to Eq. 4 with $r_{ab}^{\text{empirical}} \propto e^{\beta w_{ab}^{\text{estimate}}}$.

In the present analyses, we assume for simplicity that $m_{i,\mu_i\nu_i}$ and $f_{i,\nu_i}^{\text{ mut}}$ do not depend on codon position i; that is, $m_{i,\xi\eta} = m_{\xi\eta}$ and $f_{i,\xi}^{\text{ mut}} = f_{\xi}^{\text{ mut}}$, where $\xi,\eta \in \{a,t,c,g\}$. This approximation is reasonable because mutational tendencies may be independent of a nucleotide position in a codon. Let us define $m_{[tc][ag]}$ to represent the average of the exchangeabilities of the transversion type, m_{ta}, m_{tg}, m_{ca}, and m_{cg}, and likewise $m_{tc|ag}$ to represent the average of the exchangeabilities of the transition type, m_{tc} and m_{ag}. We use the ratios $\{m_{\xi\eta}/m_{[tc][ag]}\}$ as parameters for exchangeabilities, and $m(\equiv m_{[tc][ag]})$ to represent the ratio of the exchangeability of double nucleotide change to that of single nucleotide change and also the ratio of the exchangeability of triple nucleotide change to that of double nucleotide change; note that the exchangeabilities of single, double, and triple nucleotide changes are of $O(m_{[tc][ag]}), O(m_{[tc][ag]}^2)$, and $O(m_{[tc][ag]}^3)$ in Eq. 3, respectively, and that Eq. 3 must be satisfied. Then, multiple nucleotide changes in infinitesimal time can be completely neglected by making the parameter $m(\equiv m_{[tc][ag]})$ approach zero with keeping $\{m_{\xi\eta}/m_{[tc][ag]}\}$ constant in Eq. 3. Also, it is noted that unlike the SDT model [20] double nucleotide changes at the first and the third positions in a codon are assumed to occur as frequently as doublet changes.

The number of parameters except equilibrium codon frequencies in the mechanistic codon substitution model is equal to 11; they are β, w_0, $m(\equiv m_{[tc][ag]})$, $m_{tc|ag}/m_{[tc][ag]}$, $m_{ag}/m_{tc|ag}$, $m_{ta}/m_{[tc][ag]}$, $m_{tg}/m_{[tc][ag]}$, $m_{ca}/m_{[tc][ag]}$, $f_a^{\text{ mut}}$, $f_c^{\text{ mut}}$, and $f_g^{\text{ mut}}$, and fixed at certain values or optimized as ML parameters.

Variations of mutation rate and of selective constraint across codon sites

Taking account of the variation of amino acid substitution rate over sites always increases the maximum likelihood of a phylogenetic tree in the analysis of amino acid sequences [31]. The variation of amino acid substitution rate can be caused by the variation of mutation rate and also by the variation of selective constraint on amino acids. Here, the variation of either mutation rate or selective constraint over sites is taken into account, but both are not taken into account at the same time because of a heavy computational load.

The variation of mutation rate over codon sites is also assumed to obey a Γ distribution [31] with a shape parameter α and the mean equal to 1, which is then approximated by a discrete-gamma distribution [32,40] with m categories, each with equal probability, This model is specified with a suffix dGmr whose m means the number of categories.

The variation of selective constraint over amino acid sites is assumed to obey a discrete-gamma distribution, too. In this model, the average of selective constraints over amino acid pairs (the mean acceptance rate), $\sum_a \sum_{b>a} e^{w}{}_{ab}/190$ in the mechanistic codon substitution model or w_0 in other codon substitution models, is assumed to vary according to a discrete-gamma distribution. The rate matrix of each category is scaled so that the mean rate matrix satisfies Eq. 3. This model is specified with a suffix dGms whose m means the number of categories.

In the mechanistic codon substitution model, selective constraint $w_{i,ab}$ for ith category in a discrete-gamma distribution is calculated to satisfy the following equations.

$$\sum_i \Gamma_i p(\Gamma_i) = \frac{1}{190} \sum_a \sum_{b>a} e^{w}{}_{ab} \tag{14}$$

$$\frac{1}{190} \sum_a \sum_{b>a} e^{w_{i,ab}} = \Gamma_i \tag{15}$$

$$e^{w_{i,ab}} \equiv \begin{cases} \min[\gamma_i \exp(\beta w_{ab}^{\text{estimate}} + w_0(1-\delta_{ab})),1] & \text{for } \Gamma_i < 1 \\ \Gamma_i & \text{for } \Gamma_i \geq 1 \end{cases} \tag{16}$$

where Γ_i is the value of the ith category in the discrete-gamma distribution whose mean is equal to the average of $e^{w_{ab}}$ over all amino acid pairs and whose shape parameter is equal to α; $0 \leq \Gamma_i < \Gamma_{i+1}$. If $\Gamma_i < 1$ and $\gamma_i \exp w_{ab} \leq 1$ *for* $\forall a,b$, γ_i will be simply equal to a point of the discrete-gamma distribution whose mean is equal to 1.

In the other codon models, the equal amino acid constraint $w_{i,0}$ for ith category in Eq. 4 and Eq. 5 is calculated from the following equations.

$$\sum_i \Gamma_i p(\Gamma_i) = e^{w0} \tag{17}$$

$$e^{w_{i,0}} = \Gamma_i = \gamma_i e^{w0} \tag{18}$$

In this case, Γ_i is and γ_i are points of the discrete-gamma distributions, whose means are equal to $\exp w_0$ and 1, respectively, with the shape parameter α.

The shape parameter α of the discrete-gamma distribution for the variation of mutation rate or selective constraint is optimized as one of ML parameters. Equal probability of each category is used for the mutation rate variation, but it may be inappropriate for the variation of selective constraint, because $\Gamma_i (i>1)$ is often too small for a rate matrix to be significantly different between Γ_{i-1} and Γ_i. In such a case, the prior probability of Γ_{i-1} is increased to make the rate matrices for Γ_{i-1} and Γ_i significantly different.

A simple approximation for the variation of mutation rate over time

A mutation rate at each site may vary in each branch, especially long branches, of a phylogenetic tree. If the variation of mutation rate is synchronized among sites, it will be reflected by the length of each branch. The unsynchronized portion of rate variation among sites is considered. Here, a simple approximation for the variation of mutation rate over time is provided. The mutation rate matrix M and therefore the substitution rate matrix R are assumed to vary in time only by a scalar factor, $\mu(t)$ at time t. The expected values of the mean and the variance of the total substitution rate in a branch whose length is equal to T are as follows.

$$E(\int_0^T \mu(t)dt) = E(\mu)T \tag{19}$$

$$E((\int_0^T (\mu(t)-E(\mu))dt)^2) = \int_0^T \int_0^T E((\mu(t)-E(\mu))(\mu(t')-E(\mu)))dtdt'$$

$$\simeq 2\tau E((\mu - E(\mu))^2)T \quad (\tau \ll T)) \tag{20}$$

The mutation rate as a function of time is assumed to be autocorrelated with a correlation time $\tau \ll T$. In this case, the mean and the variance are both the linear functions of T. For the variation of the total mutation rate in the branch of the length T, we assume a Γ distribution whose scale and shape parameters are equal to $\sigma \simeq 2\tau E((\mu - E(\mu))^2)/E(\mu)$ and $\alpha = E(\mu)T/\sigma$, respectively. Then, the expected substitution matrix is:

$$E(S(E(\mu),T)) \equiv \int_0^\infty e^{Rx} \Gamma(x; E(\mu)T/\sigma, \sigma)dx$$

$$= \int_0^\infty \frac{1}{\Gamma(E(\mu)T/\sigma)} \exp\{-(I-\sigma R)\frac{x}{\sigma}\}(\frac{x}{\sigma})^{E(\mu)T/\sigma - 1} \frac{dx}{\sigma}$$

$$= (I - \sigma R)^{-E(\mu)T/\sigma} = (E(S(1,1)))^{E(\mu)T} \tag{21}$$

where $\Gamma(x; E(\mu)T/\sigma, \sigma)$ is the probability density function of a Γ distribution with a scale parameter σ and a shape parameter equal to $E(\mu)T/\sigma$, $\Gamma(E(\mu)T/\sigma)$ is the Γ function, and I is the identity matrix. Then, $\log E(S(1,1))$ is used instead of R as a rate matrix; the rate matrix $\log E(S(1,1))$ is scaled to make the mean rate matrix satisfy Eq. 3. A constant mutation rate corresponds to $\sigma = 0$. The scale parameter σ is set to 0 or is optimized as a ML parameter.

This approximation for the variation of mutation rate over time is very simple and does not require any additional computational time, although the performance will be limited in comparison with a more complete approximation [36]. However, the ML estimate of σ in this approximation may be influenced by the variation of mutation rate across sites, because the mean of the substitution matrix over sites is represented by a similar functional form to $E(S(E(\mu),T))$; assuming that mutation rates vary across sites with a Γ distribution, the mean of substitution matrix over sites for a branch of the length T is formulated as $\langle S(E(\mu),T) \rangle \equiv \int_0^\infty \exp(RTx)\Gamma(x; E(\mu)/\sigma, \sigma)dx = (I - \sigma RT)^{-E(\mu)/\sigma}$, which is equal to the expected substitution matrix in the case of $\tau > T$.

Datasets of protein-coding sequences used to evaluate codon substitution models

Substitution models are evaluated by using five datasets of codon sequences; (1) divergent and (2) closely-related chloroplast-encoded genes, (3) fast-evolving interspecific and (4) highly-polymorphic intraspecific mitochondrial genes, and (5) slowly-evolving nuclear genes.

1. Dataset cpDNA-9: Divergent codon sequences consisting of 45 protein-coding genes from 9 chloroplast genomes, whose protein sequences were used to estimate the cpREV10 by [8]; *Synechocystis PCC6803*, which was the outgroup sequence in their analysis, is not used in the present analysis. The codon sequences were obtained from the NCBI RefSeq database of organelle genomes. The total codon length of aligned genes is equal to 12507, and the minimum amino acid identity between sequences is equal to 0.58. The tree topology that was estimated as Tree-1 by [8] is used here as the most probable tree. Overlapped segments between genes were removed from codon sequences.
2. Dataset cpDNA-55: Codon sequences consisting of 52 protein-coding genes from 55 chloroplast genomes of the major angiosperm lineages, which are genome sequences available in the NCBI RefSeq database out of the 64 genomes analyzed in [41], and which are genes owned by all 55 taxa. The tree topology estimated by [41] is used as the most probable tree in the present analysis. The total codon length of aligned genes is equal to 14128, and the minimum amino acid identity between the sequences is equal to 0.73. The cpREV64 [38] was estimated from the full set of 77 protein-coding genes in the 64 genomes.
3. Dataset mammalian-mtDNA: Interspecific mammalian mitochondrial codon sequences consisting of 12 protein-coding genes from 69 mammalian species [42], whose genome sequences were obtained from the NCBI RefSeq database of organelle genomes. The total codon length of aligned genes is equal to 3618, and the minimum amino acid identity between the sequences is equal to 0.66. The tree topology that was estimated as Tree-6 by [42] is used here as the most probable tree. Overlapped segments between genes were removed from codon sequences.
4. Dataset human-mtDNA: Intraspecific human mitochondrial codon sequences consisting of 12 protein-coding genes from 53 human races [43], whose genome sequences were obtained from a human mitochondrial genome database (MITOMAP). The total codon length of aligned genes is equal to 3579, and the minimum amino acid identity between the sequences is equal to 0.99. The present analyses are done using the neighbor-joining tree topology estimated by [43]. Overlapped segments between genes were removed from codon sequences.
5. Dataset nDNA: Codon sequences of the 10 most slowly-evolving genes out of the 2789 nuclear genes of 10 mammals that were analyzed by [44]. The tree topologies estimated by [44] are used for respective genes and the tree-1 named by them is used here for the analyses of the concatenated genes. The total codon length of aligned genes is equal to 1112, and the minimum amino acid identity between the sequences is equal to 0.97.

Homologous codon sequences are aligned every gene by ClustalW2 [45] that is modified to align codon sequences with codon score matrices [19]. The ML values for each model are calculated for each gene and also for the concatenated sequences of all genes by Phyml [46] also modified to analyze codon sequences.

Statistical comparison of codon substitution models

Model selection must be pursued with considerable attention [47]. For the comparison of models one of which is a special case of the other, the likelihood ratio test (LRT) [48] can be used to test the superiority of a nesting model to nested models. Models that are not nesting or nested can be compared using Akaike information criterion (AIC) [49], Bayesian information criterion (BIC) [50], a decision-theoretical approach [51,52], and the Bayes factor [53]. Here, AIC and BIC for a given tree topology of aligned codon sequences are used to compare codon substitution models derived from various empirical amino acid and codon substitution rate matrices and mechanistic codon substitution models with the wide range of selective constraint matrices. The AIC and BIC are defined as follows [18]:

$$\mathrm{AIC} \equiv -2\ell(\hat{\boldsymbol{\theta}}) + 2K \tag{22}$$

$$\mathrm{BIC} \equiv -2\ell(\hat{\boldsymbol{\theta}}) + K \log n \tag{23}$$

where K is the number of adjustable parameters, $\hat{\boldsymbol{\theta}}$ is the vector of the ML estimates of the parameters, $\ell(\hat{\boldsymbol{\theta}})$ is the maximum log-likelihood value, and n is the number of codons in a codon alignment. The model whose AIC or BIC is the minimum is regarded as the best model.

Results and Discussion

The naming convention of the present models is briefly described in Table 1. In all models, the equilibrium frequencies of codons are estimated to be equal to codon frequencies in sequences. Other parameters including the scale parameter σ of a Γ distribution for the variation of mutation rate over time are set to a certain value or optimized by maximizing the likelihood of a given topology of a phylogenetic tree. For the empirical amino acid substitution models converted into the codon substitution models, $\sigma = 0$ was assumed, because it seems not to be well matched with these models. In the empirical codon substitution model, σ was optimized as well as w_0. In the mechanistic codon substitution models, all 12 parameters including σ for the substitution rate matrix will be optimized if the AIC and the BIC values of a phylogenetic tree are decreased. For the separating analyses of human-mtDNA and the concatenating and the separating analyses of nDNA, which are both datasets consisting of highly-homologous sequences, the five parameters of β, w_0, σ, $m(\equiv m_{[tc][ag]})$, and $m_{tc|ag}/m_{[tc][ag]}$ were optimized with $f^{\,\mathrm{mut}_{\xi}} = 0.25$ and $m_{ag}/m_{tc|ag} = m_{ta}/m_{[tc][ag]} = m_{tg}/m_{[tc][ag]} = m_{ca}/m_{[tc][ag]} = 1.0$. In all models, the variation of mutation rate or the variation of selective constraint over sites is taken into account. Both the variations over sites were approximated by a discrete-gamma distribution [32] with 4 categories. The shape parameter α of the discrete-gamma distribution is optimized by maximizing the likelihood. Equal probability was used for each category in all models of rate variation. In the models of variable selective constraints, equal probability was used only for the non-mechanistic codon models for the cpDNA-9, and the different sets of prior probabilities on the basis of the values of Γ_i were used for the other models; $p(\Gamma_1) = 0.50, p(\Gamma_2) = 0.25, p(\Gamma_3) = p(\Gamma_4) = 0.125$ for the datasets cpDNA-9, cpDNA-55, and mammalian-mtDNA, and $p(\Gamma_1) = 0.75, p(\Gamma_2) = 0.125, p(\Gamma_3) = p(\Gamma_4) = 0.0625$

Table 1. Brief description of models.

A. Empirical amino acid substitution models converted into codon substitution models	
JTT-*n* -F-dG*m*[rs],[a] WAG-*n* -F-dG*m*[rs], LG-*n* -F-dG*m*[rs], cpREV10-*n* -F-dG*m*[rs], cpREV64-*n* -F-dG*m*[rs], mtREV-*n* -F-dG*m*[rs]	The empirical amino acid exchangeabilities of JTT [5], WAG [10], LG [11], cpREV10 [8], cpREV64 [38], and mtREV [6] are used as $\{r_{ab}^{\text{empirical}}\}$ in Eq. 4. The suffix *n* means the number of parameters optimized for the substitution rate matrix; the w_0 is a ML parameter when $n \geq 1$.
B. Empirical codon substitution models	
KHG-*n* -F-dG*m*[rs][a]	The empirical codon exchangeabilities of KHG [15] are used as $\{r_{\mu\nu}^{\text{empirical}}\}$ in Eq. 5. The suffix *n* means the number of parameters optimized for the substitution rate matrix; the w_0 is a ML parameter when $n \geq 1$, and σ is equal to 0 for $n=1$ and optimized when $n=2$.
C. Mechanistic codon substitution models	
No-Constraint-*n* -F-dG*m*r[a], Equal-Constraint-*n* -F-dG*m*[rs]	$\beta=0$ for both models and also $w_0=0$ for the No-Constraint model; see Eq. 9. The suffix *n*, whose maximum number is equal to 10 or 11, means the number of parameters optimized for the substitution rate matrix.
EI-*n* -F-dG*m*[rs][a]	$\hat{w}_{ab}^{\text{estimate}} \equiv \Delta\hat{\varepsilon}_{ab}^{c} + \Delta\hat{\varepsilon}_{ab}^{v}$ based on the Energy-Increment-based (EI) method [19] is used to estimate w_{ab} in Eq. 9. The $\Delta\hat{\varepsilon}_{ab}^{c}$ and $\Delta\hat{\varepsilon}_{ab}^{v}$ represent the mean increment of contact energies between residues, and the mean volume change due to an amino acid replacement, respectively; see Supporting Information, Text S1, in Miyazawa [19]. The suffix *n*, whose maximum number is equal to 12, means the number of parameters optimized for the substitution rate matrix.
JTT-ML91+-*n* -F-dG*m*[rs],[a] WAG-ML91+-*n* -F-dG*m*[rs], LG-ML91+-*n* -F-dG*m*[rs]	Selective constraints $\{w_{ab}^{\text{JTT/WAG/LG}-\text{ML91}+}\}$ estimated by maximizing the likelihood of JTT/WAG/LG [5,10,11] in the ML-91+ model [19] are used as $\{w_{ab}^{\text{estimate}}\}$ in Eq. 9. The suffix *n*, whose maximum number is equal to 12, means the number of parameters optimized for the substitution rate matrix.
KHG-ML200-*n* -F-dG*m*[rs][a]	Selective constraints $\{w_{ab}^{\text{KHG}-\text{ML200}}\}$ estimated by maximizing the likelihood of the KHG codon substitution matrix [15] in the ML-200 model [19] are used as $\{w_{ab}^{\text{estimate}}\}$ in Eq. 9. The suffix *n*, whose maximum number is equal to 12, means the number of parameters optimized for the substitution rate matrix.

[a] In the models specified with the suffix "F", equilibrium codon frequencies are assumed to be equal to codon frequencies in codon sequences. dG*m*[rs], i.e., dG*m*r or dG*m*s, means that the variation of mutation rate or selective constraint over site is approximated by a discrete gamma distribution with *m* categories [32], respectively; $m=1$ means no variation and the suffix dG*1*[rs] is omitted.

doi:10.1371/journal.pone.0028892.t001

for the human-mtDNA and the nDNA. The AIC and the BIC are used for statistical comparisons of models.

Mechanistic codon substitution models outperform other substitution models

First, each gene in a dataset is separately aligned and then all aligned sequences are concatenated. The maximum log-likelihood values of a given phylogenetic tree of concatenated genes for various codon substitution models are listed in Tables 2, 3, 4, 5, 6 for cpDNA-9, cpDNA-55, mammalian-mtDNA, human-mtDNA, and nDNA, respectively. Values in parentheses indicate that the corresponding parameters are fixed at the value specified. The maximum log-likelihood (ℓ), AIC and BIC values for each model are listed in these tables with the difference ($\Delta\ell$, $\Delta \mathrm{AIC}$, and $\Delta \mathrm{BIC}$) from those of a reference model. For the datasets cpDNA-9, cpDNA-55, and nDNA that use the universal codon table, the empirical codon substitution model KHG-2-F-dG4s estimated from nuclear-encoded sequences is used as a reference state; in the KHG-2-F-dG4s, σ is optimized as well as w_0. For mitochondrial genomes that use a minor genetic code, no empirical codon substitution rate matrix is available, and so the codon substitution model, mtREV-1-dG4s, which is converted from the empirical amino acid substitution matrix mtREV estimated from mitochondrial proteins, is used as a reference state; in the mtREV-1-F-dG4s, $\sigma=0$ is assumed, and only w_0 is optimized.

In the case of mitochondrial genes, i.e., mammalian-mtDNA and human-mtDNA, the models based on mtREV always show the smallest $\Delta \mathrm{AIC}$ and $\Delta \mathrm{BIC}$, i.e., the best performance, in the empirical amino acid substitution models converted into the codon substitution models. For the dataset cpDNA-55, the models converted from cpREV64 show the best performance in the models converted from the empirical amino acid substitution models, and the models converted from cpREV10 perform best for the dataset cpDNA-9. These results are reasonable because the amino acid substitution probability matrix mtREV [6] was estimated from mitochondrial proteins, and cpREV64 [38] and cpREV10 [8] were estimated from the full sets of chloroplast proteins corresponding to cpDNA-55 and cpDNA-9, respectively; see the method section. A rather interesting result is that the models converted from cpREV64 shows larger $\Delta \mathrm{AIC}$ and $\Delta \mathrm{BIC}$ for cpDNA-9 than the models converted from LG, WAG, and JTT that were estimated from nuclear-encoded proteins, This fact indicates that substitution tendencies vary between genes and cannot always be represented by the average tendencies of substitutions. Delport et al.[16] showed that the empirical substitution matrices represent the average tendencies of substitutions over various protein families by sacrificing gene-level resolution.

The empirical codon substitution model KHG performs significantly better for chloroplast-encoded and nuclear-encoded genes than all the amino acid substitution models converted into the codon models. It has often be insisted that synonymous substitutions are saturated between distantly related genes and so substitution analyses at the codon level hardly include more information than those at the amino acid level. However, a fact that KHG performs better even for the distantly related sequence family (cpDNA-9) than the models converted from cpREV10 indicates that codon sequences include more information than amino acid sequences even in the case of distantly related sequences.

If the amino acid substitution models converted into codon models are compared with the mechanistic codon substitution models, the superiority of the codon substitution models will be clearer. For all datasets, the mechanistic codon models with the various estimates of selective constraints show significantly lower

Table 2. Comparisons between various codon substitution models in the concatenating analysis of cpDNA-9.

Codon substitution model[a]	**K**[b]	$\Delta\ell$[c]	ΔAIC[c]	ΔBIC[c]	$\langle e^{w_{ab}} \rangle$[de]	$\hat{\sigma}$[ef]	$\hat{m}$[eg]	[h]	$\hat{\alpha}$[i]
cpREV64-1-F-dG4r	62	−3180.3	6358.6	6351.2		(0.0)			0.977
LG-1-F-dG4r	62	−2912.8	5823.6	5816.1		(0.0)			0.973
JTT-1-F-dG4r	62	−2608.4	5214.8	5207.4		(0.0)			1.020
WAG-1-F-dG4r	62	−2501.2	5000.5	4993.0		(0.0)			1.148
cpREV10-1-F-dG4r	62	−1575.7	3149.3	3141.9		(0.0)			1.195
cpREV64-1-F-dG4s	62	−1504.2	3006.4	2999.0		(0.0)			0.505
LG-1-F-dG4s	62	−1321.4	2640.8	2633.4		(0.0)			0.496
WAG-1-F-dG4s	62	−1126.3	2250.5	2243.1		(0.0)			0.573
JTT-1-F-dG4s	62	−1046.0	2090.0	2082.6		(0.0)			0.519
cpREV10-1-F-dG4s	62	−284.0	566.0	558.6		(0.0)			0.591
KHG-2-F-dG4r	63	−1237.8	2475.7	2475.7		0.031			1.301
KHG-2-F-dG4s	63	0.0	0.0	0.0		0.290			0.575
No-Constraint-10-F-dG4r	71	−19392.6	38801.2	38860.7	(1.0)	0.000	0.040	2.541	1.830
Equal-Constraint-11-F-dG4r	72	−1355.6	2729.2	2796.1	0.021	0.424	0.292	2.053	1.178
EI-12-F-dG4r	73	−253.1	526.3	600.6	0.023	0.000	0.494	2.217	1.160
JTT-ML91+-12-F-dG4r	73	288.7	−557.3	−483.0	0.018	0.002	0.569	1.702	1.131
WAG-ML91+-12-F-dG4r	73	477.4	−934.7	−860.4	0.015	0.272	0.526	2.184	1.126
KHG-ML200-12-F-dG4r	73	562.9	−1105.8	−1031.5	0.039	0.000	0.325	1.610	1.122
LG-ML91+-12-F-dG4r	73	627.3	−1234.6	−1160.3	0.023	0.000	0.485	2.158	1.144
Equal-Constraint-11-F-dG4s	72	680.2	−1342.4	−1275.5	0.063	0.414	0.208	2.196	0.384
EI-12-F-dG4s	73	1935.2	−3850.4	−3776.0	0.060	0.000	0.431	2.307	0.390
JTT-ML91+-12-F-dG4s	73	2640.2	−5260.4	−5186.1	0.052	0.125	0.461	1.774	0.363
KHG-ML200-12-F-dG4s	73	2646.5	−5273.0	−5198.6	0.106	0.170	0.215	1.705	0.388
WAG-ML91+-12-F-dG4s	73	2827.2	−5634.4	−5560.1	0.048	0.313	0.405	2.349	0.359
LG-ML91+-12-F-dG4s	73	2956.6	−5893.1	−5818.8	0.064	0.201	0.364	2.369	0.370
LG-ML91+-11s-F-dG4s	72	2412.8	−4807.6	−4740.6	0.066	2.335	(0.0)	2.667	0.297
LG-ML91+-11-F-dG4s	72	2942.4	−5866.8	−5799.9	0.066	(0.0)	0.409	2.292	0.385
LG-ML91+-12-F	72	−1833.2	3684.4	3751.4	0.026	0.878	0.622	2.039	

[a] The prior probability of each category for the mechanistic codon models of "dG4s" is $p(\Gamma_1)=0.50$, $p(\Gamma_2)=0.25$, and $p(\Gamma_3)=p(\Gamma_4)=0.125$; equal probability is used in other models.

[b] The number of adjustable parameters.

[c] Differences from the reference state; $\Delta\ell=\ell+200811.7$, $\Delta\mathrm{AIC}=(-2\ell+2K)-401749.5$, and $\Delta\mathrm{BIC}=(-2\ell+K\log 12507)-402217.8$.

[d] The average of e^{w}_{ab} over all amino acid pairs $\{a,b\}$; $\langle e^{w_{ab}} \rangle \equiv \frac{1}{190}\sum_a \sum_{b>a} e^{w_{ab}}$.

[e] The value parenthesized means that the parameter is fixed at the value specified.

[f] The scale parameter of a Γ distribution for the variation of mutation rate over time.

[g] The ratio of double to single and of triple to double nucleotide change exchangeability; $\hat{m} \equiv \hat{m}_{[tc][ag]}$.

[h] The ratio of mean transitional to mean transversional exchangeability; $\hat{m}_{tc|ag}/\hat{m}_{[tc][ag]}$.

[i] The shape parameter of a discrete gamma distribution for the variation of mutation rate or selective constraint over sites.

doi:10.1371/journal.pone.0028892.t002

ΔAIC and ΔBIC than the amino acid substitution models converted into the codon models. The Equal-Constraint model always performs worst, and is far inferior to the amino acid dependent constraint models for the phylogenetic trees including long branches such as the datasets cpDNA-9 and mammalian-mtDNA. Only for the phylogenetic trees consisting of extremely short branches such as the datasets human-mtDNA and nDNA, it is not remarkably worse than the amino acid dependent constraint models; amino acid identities between sequences are equal to or larger than 0.99 in human-mtDNA and 0.97 in nDNA. Consistently, ΔAIC and ΔBIC for the No-Constraint model, which is essentially equivalent to a nucleotide substitution model, are extremely larger for cpDNA-9 and mammalian-mtDNA, but smaller for cpDNA-55 and human-mtDNA than those for the reference model. These results can be explained to be because the amino acid dependencies of selective constraints must be taken into account to correctly evaluate amino acid substitutions, which occur in long branches, in order to precisely estimate branch lengths. One of the interesting facts is that the No-Constraint model is better for cpDNA-55 and human-mtDNA but worse for nDNA than the reference model, even though the phylogenetic tree of nDNA consists of short branches. This characteristic feature results from a fact that the genes in nDNA are slowly-evolving genes with strong selective constraints on amino acids; note that sequences in the dataset nDNA are highly homologous with amino acid identities greater than 0.97 but are collected from a wide range of mammalian species, i.e., *Borentheria*, *Xenarthra*, and *Afrotheria*.

The EI model, in which the selective constraints were evaluated on the basis of average contact energies between residues in

Table 3. Comparisons between various codon substitution models in the concatenating analysis of cpDNA-55.

Codon substitution model[a]	**K**[b]	$\Delta\ell$[c]	ΔAIC[c]	ΔBIC[c]	$\langle e^{w_{ab}}\rangle$[de]	$\hat{\sigma}$[ef]	$\hat{m}$[eg]	[h]	$\hat{\alpha}$[i]
LG-1-F-dG4r	62	-15686.1	31370.2	31362.6		(0.0)			1.055
WAG-1-F-dG4r	62	-13111.9	26221.8	26214.3		(0.0)			1.094
cpREV10-1-F-dG4r	62	-11200.7	22399.4	22391.8		(0.0)			1.096
JTT-1-F-dG4r	62	-10457.8	20913.6	20906.1		(0.0)			1.092
cpREV64-1-F-dG4r	62	-6897.0	13792.0	13784.5		(0.0)			1.091
LG-1-F-dG4s	62	-11104.1	22206.2	22198.7		(0.0)			0.289
WAG-1-F-dG4s	62	-8713.6	17425.2	17417.7		(0.0)			0.300
cpREV10-1-F-dG4s	62	-6712.3	13422.7	13415.1		(0.0)			0.298
JTT-1-F-dG4s	62	-5820.8	11639.7	11632.1		(0.0)			0.299
cpREV64-1-F-dG4s	62	-1958.7	3915.4	3907.8		(0.0)			0.299
KHG-2-F-dG4r	63	-3161.2	6322.5	6322.5		0.068			1.073
KHG-2-F-dG4s	63	0.0	0.0	0.0		0.150			0.277
No-Constraint-10-F-dG4r	71	1705.9	-3395.8	-3335.3	(1.0)	0.000	0.018	3.557	1.055
Equal-Constraint-11-F-dG4r	72	26281.9	-52545.9	-52477.9	0.156	0.000	0.101	2.671	1.107
EI-12-F-dG4r	73	26941.5	-53863.0	-53787.5	0.143	0.000	0.107	2.732	1.100
JTT-ML91+-12-F-dG4r	73	27198.5	-54377.0	-54301.4	0.122	0.000	0.122	2.501	1.111
WAG-ML91+-12-F-dG4r	73	27378.4	-54736.7	-54661.2	0.125	0.000	0.115	2.690	1.100
LG-ML91+-12-F-dG4r	73	27664.8	-55309.7	-55234.1	0.142	0.000	0.112	2.707	1.109
KHG-ML200-12-F-dG4r	73	27683.4	-55346.8	-55271.2	0.163	0.000	0.099	2.479	1.106
Equal-Constraint-11-F-dG4s	72	34659.7	-69301.4	-69233.4	0.276	0.124	0.056	2.664	0.259
EI-12-F-dG4s	73	35716.3	-71412.7	-71337.1	0.235	0.103	0.071	2.727	0.247
KHG-ML200-12-F-dG4s	73	36243.5	-72467.0	-72391.4	0.251	0.116	0.058	2.477	0.285
JTT-ML91+-12-F-dG4s	73	36257.9	-72495.7	-72420.2	0.204	0.072	0.098	2.438	0.231
WAG-ML91+-12-F-dG4s	73	36362.6	-72705.2	-72629.6	0.222	0.109	0.074	2.670	0.234
LG-ML91+-12-F-dG4s	73	36583.3	-73146.6	-73071.1	0.233	0.105	0.073	2.701	0.256
LG-ML91+-11s-F-dG4s	72	36336.9	-72655.9	-72587.9	0.250	0.260	(0.0)	2.788	0.237
LG-ML91+-11-F-dG4s	72	36479.9	-72941.8	-72873.7	0.213	(0.0)	0.123	2.623	0.273
LG-ML91+-12-F	72	14390.7	-28763.5	-28695.5	0.135	0.000	0.182	2.569	

[a] The prior probability of each category for the "dG4s" is $p(\Gamma_1)=0.50$, $p(\Gamma_2)=0.25$, and $p(\Gamma_3)=p(\Gamma_4)=0.125$.
[b] The number of adjustable parameters.
[c] Differences from the reference state; $\Delta\ell=\ell+490663.8$, $\Delta\mathrm{AIC}=(-2\ell+2K)-981453.6$, and $\Delta\mathrm{BIC}=(-2\ell+K\log 14128)-981929.6$.
[d] The average of e^{w}_{ab} over all amino acid pairs $\{a,b\}$; $\langle e^{w_{ab}}\rangle\equiv\frac{1}{190}\sum_a\sum_{b>a} e^{w_{ab}}$.
[e] The value parenthesized means that the parameter is fixed at the value specified.
[f] The scale parameter of a Γ distribution for the variation of mutation rate over time.
[g] The ratio of double to single and of triple to double nucleotide change exchangeability; $\hat{m}\equiv\hat{m}_{[tc][ag]}$.
[h] The ratio of mean transitional to mean transversional exchangeability; $\hat{m}_{tc|ag}/\hat{m}_{[tc][ag]}$.
[i] The shape parameter of a discrete gamma distribution for the variation of mutation rate or selective constraint over sites.

doi:10.1371/journal.pone.0028892.t003

proteins [19], always show better performance than the Equal-Constraint model but is always inferior to the other models, which use the selective constraints estimated from the empirical amino acid substitution frequency matrices, especially for the datasets cpDNA-9 and mammalian-mtDNA including long branches. The similar result was obtained in [19]. The selective constraint matrix LG-ML91+ performs better on average than the WAG-ML91+, JTT-ML91+, and KHG-ML200, although the differences of $\Delta\ell$ between them are small in comparison with the differences from the EI. An unexpected fact is that the selective constraint matrix KHG-ML200 estimated from the codon substitution rate matrix KHG tends to be inferior to the other selective constraint matrices estimated from the empirical amino acid substitution rate matrices, LG-ML91+, WAG-ML91+, and LG-ML91+, although it performs better except for nDNA than the EI.

In the concatenating analyses of multiple genes, it is assumed that all genes have no difference in equilibrium codon frequencies, nucleotide exchangeabilities, and the variations of mutation rate and of selective constraint. These assumptions are not always appropriate. Thus, the separating analyses of multiple genes have been carried out. The ΔBIC of each gene for some models are plotted against the maximum log-likelihood value for the best model in Fig. 1 for all datasets. In all datasets, the mechanistic codon substitution models show significantly lower ΔBIC than the best amino acid substitution model converted into the codon models, for almost all genes except some genes for which the maximum log-likelihood values are large owing to short sequences. The No-Constraint model is not shown for cpDNA-9 and mammalian-mtDNA, because its ΔBIC values for them are too large to show. For the phylogenetic trees of cpDNA-55 and

Table 4. Comparisons between various codon substitution models in the concatenating analysis of mammalian-mtDNA.

Codon substitution model[a]	**K**[b]	$\Delta\ell$[c]	$\Delta\mathrm{AIC}$[c]	$\Delta\mathrm{BIC}$[c]	$\langle e^{w_{ab}} \rangle$[de]	$\hat{\sigma}$[ef]	$\hat{m}$[eg]	[h]	$\hat{\alpha}$[i]
WAG-1-F-dG4r	62	−5227.0	10454.0	10454.0		(0.0)			0.786
LG-1-F-dG4r	62	−5154.9	10309.8	10309.8		(0.0)			0.771
JTT-1-F-dG4r	62	−3673.9	7347.7	7347.7		(0.0)			0.783
mtREV-1-F-dG4r	62	−1863.0	3725.9	3725.9		(0.0)			0.870
WAG-1-F-dG4s	62	−2662.6	5325.2	5325.2		(0.0)			0.327
LG-1-F-dG4s	62	−2628.1	5256.1	5256.1		(0.0)			0.314
JTT-1-F-dG4s	62	−1285.1	2570.2	2570.2		(0.0)			0.329
mtREV-1-F-dG4s	62	0.0	0.0	0.0		(0.0)			0.339
No-Constraint-10-F-dG4r	71	−63614.9	127247.9	127303.6	(1.0)	0.000	0.000	4.908	1.965
Equal-Constraint-11-F-dG4r	72	464.7	−909.4	−847.5	0.013	0.000	0.108	4.508	0.495
EI-12-F-dG4r	73	4336.4	−8650.8	−8582.7	0.007	0.000	0.271	4.697	0.928
KHG-ML200-12-F-dG4r	73	5340.3	−10658.5	−10590.4	0.022	0.000	0.088	3.238	0.480
JTT-ML91+-12-F-dG4r	73	5501.7	−10981.4	−10913.3	0.006	0.000	0.228	3.679	0.452
WAG-ML91+-12-F-dG4r	73	5728.6	−11435.1	−11367.0	0.006	0.000	0.206	5.614	0.492
LG-ML91+-12-F-dG4r	73	6315.1	−12608.2	−12540.0	0.009	0.000	0.147	5.921	0.515
Equal-Constraint-11-F-dG4s	72	6961.9	−13903.8	−13841.9	0.036	1.313	0.031	4.984	0.269
EI-12-F-dG4s	73	10402.4	−20782.8	−20714.6	0.024	1.137	0.124	5.426	0.267
KHG-ML200-12-F-dG4s	73	11621.0	−23219.9	−23151.8	0.063	1.119	0.039	3.658	0.306
JTT-ML91+-12-F-dG4s	73	11698.3	−23374.5	−23306.4	0.022	1.637	0.091	4.189	0.259
WAG-ML91+-12-F-dG4s	73	11997.4	−23972.8	−23904.6	0.020	1.686	0.092	6.588	0.259
LG-ML91+-12-F-dG4s	73	12532.5	−25042.9	−24974.8	0.028	1.826	0.065	7.158	0.262
LG-ML91+-11-F-dG4s	72	12113.1	−24206.3	−24144.4	0.035	(0.0)	0.128	6.009	0.290
LG-ML91+-11s-F-dG4s	72	12268.3	−24516.5	−24454.6	0.028	3.066	(0.0)	7.600	0.252
LG-ML91+-12-F	72	−4803.5	9627.1	9689.0	0.011	3.713	0.196	5.477	

[a] The prior probability of each category for the "dG4s" is $p(\Gamma_1)=0.50$, $p(\Gamma_2)=0.25$, and $p(\Gamma_3)=p(\Gamma_4)=0.125$.
[b] The number of adjustable parameters.
[c] Differences from the reference state; $\Delta\ell=\ell+343200.7$, $\Delta\mathrm{AIC}=(-2\ell+2K)-686525.4$, and $\Delta\mathrm{BIC}=(-2\ell+K\log 3618)-686909.5$.
[d] The average of e^{w}_{ab} over all amino acid pairs $\{a,b\}$; $\langle e^{w_{ab}} \rangle \equiv \frac{1}{190}\sum_a \sum_{b>a} e^{w_{ab}}$.
[e] The value parenthesized means that the parameter is fixed at the value specified.
[f] The scale parameter of a Γ distribution for the variation of mutation rate over time.
[g] The ratio of double to single and of triple to double nucleotide change exchangeability; $\hat{m}\equiv\hat{m}_{[tc][ag]}$.
[h] The ratio of mean transitional to mean transversional exchangeability; $\hat{m}_{tc|ag}/\hat{m}_{[tc][ag]}$.
[i] The shape parameter of a discrete gamma distribution for the variation of mutation rate or selective constraint over sites.

doi:10.1371/journal.pone.0028892.t004

human-mtDNA consisting of relatively short branches, the No-Constraint model, i.e., a nucleotide substitution model, is better for most of the genes than the amino acid substitution models converted into the codon models, as also indicated by the concatenating analyses. Even for those datasets, $\Delta\mathrm{BIC}$ can be further decreased by the mechanistic codon substitution models including the the Equal-Constraint model. However, differences of $\Delta\mathrm{BIC}$ between the mechanistic substitution models with the different selective constraints are small for those dataset in comparison with the improvement from the amino acid substitution models converted into the codon models. For the phylogenetic trees of cpDNA-9 and mammalian-mtDNA consisting of long branches, the differences between the Equal-Constraint and the EI and between the EI and the best model with amino acid dependent selective constraints are very significant, as indicated by the concatenating analyses.

The mechanistic codon substitution model performs better for a wide range of sequences from highly-homologous to highly-diverged sequences than both nucleotide and amino acid substitution models. This is because it takes into account both mutational tendencies at the nucleotide level and selection at the amino acid level.

Variable mutation rates versus variable selective constraints over sites

Significance of rate variation over sites in proteins has been demonstrated in nucleotide substitution models and amino acid substitution models [32,33]. These results do not necessarily indicate the variation of mutation rate over sites, because the variation of selective constraint over sites in proteins can also cause the variation of amino acid substitution rate over sites even under a uniform mutation rate over sites. Here, we examine which model better fits the heterogeneity of amino acid substitution rate over sites.

The discrete gamma distribution with 4 categories has been used to emulate both the variations of selective constraint and of mutation rate over sites. The models with variable selective constraints and with variable mutation rates are specified by dG4s and dG4r, respectively. Tables 2, 3, 4 for the concatenating analyses of genes consistently indicate that the codon substitution

Table 5. Comparisons between various codon substitution models in the concatenating analysis of human-mtDNA.

Codon substitution model[a]	K[b]	$\Delta\ell$[c]	ΔAIC[c]	ΔBIC[c]	$\langle e^{w_{ab}} \rangle$[de]	$\hat{\sigma}$[ef]	$\hat{m}$[eg]	[h]	$\hat{\alpha}$[i]
LG-1-F-dG4r	62	−42.4	84.8	84.8		(0.0)			0.338
WAG-1-F-dG4r	62	−40.1	80.1	80.1		(0.0)			0.343
JTT-1-F-dG4r	62	5.9	−11.8	−11.8		(0.0)			0.345
mtREV-1-F-dG4r	62	7.7	−15.3	−15.3		(0.0)			0.331
LG-1-F-dG4s	62	−49.2	98.3	98.3		(0.0)			0.080
WAG-1-F-dG4s	62	−46.2	92.5	92.5		(0.0)			0.079
JTT-1-F-dG4s	62	−1.8	3.7	3.7		(0.0)			0.091
mtREV-1-F-dG4s	62	0.0	0.0	0.0		(0.0)			0.085
No-Constraint-10-F-dG4r	71	315.1	−612.3	−556.6	(1.0)	0.000	0.000	47.760	0.465
Equal-Constraint-11-F-dG4r	72	515.6	−1011.2	−949.4	0.092	0.066	0.000	39.685	0.612
EI-12-F-dG4r	73	525.8	−1029.7	−961.7	0.076	0.000	0.000	36.208	0.620
KHG-ML200-12-F-dG4r	73	530.8	−1039.5	−971.5	0.110	0.000	0.000	31.182	0.648
WAG-ML91+-12-F-dG4r	73	535.5	−1049.1	−981.1	0.069	0.000	0.000	41.404	0.635
LG-ML91+-12-F-dG4r	73	535.7	−1049.4	−981.4	0.089	0.002	0.000	42.787	0.627
JTT-ML91+-12-F-dG4r	73	541.0	−1059.9	−991.9	0.051	0.000	0.000	32.733	0.646
Equal-Constraint-11-F-dG4s	72	517.4	−1014.7	−952.9	0.108	0.000	0.000	36.267	0.106
EI-12-F-dG4s	73	528.3	−1034.7	−966.7	0.079	0.000	0.000	34.994	0.123
KHG-ML200-12-F-dG4s	73	536.6	−1051.2	−983.2	0.106	0.000	0.000	30.457	0.227
LG-ML91+-12-F-dG4s	73	538.4	−1054.7	−986.7	0.078	0.000	0.000	39.024	0.233
WAG-ML91+-12-F-dG4s	73	539.3	−1056.7	−988.7	0.059	0.000	0.000	38.794	0.207
JTT-ML91+-12-F-dG4s	73	542.6	−1063.2	−995.2	0.049	0.000	0.000	32.064	0.168
JTT-ML91+-11-F-dG4s	72	542.6	−1065.2	−1003.4	0.049	(0.0)	0.000	32.067	0.168
JTT-ML91+-11s-F-dG4s	72	542.6	−1065.2	−1003.4	0.049	0.000	(0.0)	32.067	0.168
JTT-ML91+-12-F	72	522.3	−1024.6	−962.7	0.052	0.000	0.000	32.207	

[a] The prior probability of each category for the "dG4s" is $p(\Gamma_1)=0.75$, $p(\Gamma_2)=0.125$, and $p(\Gamma_3)=p(\Gamma_4)=0.0625$.

[b] The number of adjustable parameters.

[c] Differences from the reference state; $\Delta\ell=\ell+17283.0$, $\Delta\text{AIC}=(-2\ell+2K)-34690.0$, and $\Delta\text{BIC}=(-2\ell+K\log 3579)-35073.3$.

[d] The average of e^{w}_{ab} over all amino acid pairs $\{a,b\}$; $\langle e^{w_{ab}} \rangle \equiv \frac{1}{190}\sum_a \sum_{b>a} e^{w_{ab}}$.

[e] The value parenthesized means that the parameter is fixed at the value specified.

[f] The scale parameter of a Γ distribution for the variation of mutation rate over time.

[g] The ratio of double to single and of triple to double nucleotide change exchangeability; $\hat{m} \equiv \hat{m}_{[tc][ag]}$.

[h] The ratio of mean transitional to mean transversional exchangeability; $\hat{m}_{tc|ag}/\hat{m}_{[tc][ag]}$.

[i] The shape parameter of a discrete gamma distribution for the variation of mutation rate or selective constraint over sites.

doi:10.1371/journal.pone.0028892.t005

models with the variation of selective constraint (dG4s) show significantly lower ΔAIC and ΔBIC than the corresponding models with the variation of mutation rate (dG4r) over sites for the datasets cpDNA-9, cpDNA-55, and mammalian-mtDNA. The comparisons of ΔBIC of each gene between those two types of the models are shown in Fig. 2 for all datasets. These figures also show that the variation of selective constraint is a statistically better model than the variation of mutation rate at least for cpDNA-9, cpDNA-55 and mammalian-mtDNA. This is reasonable because a mutation rate may not significantly differ among sites in a gene but selective constraints originating in the tertiary structure and the function of a protein should vary among sites in a protein. Generally speaking, selective constraints on amino acid replacements are stronger in a protein core than on protein surface [54].

However, in both the concatenating analyses and the separating analyses of genes, the ΔAIC and the ΔBIC values for the models with the variation of selective constraint are not smaller for the nDNA than those for the models with the variation of mutation rate. For the human-mtDNA consisting of highly-polymorphic intraspecific mitochondrial genes, the mechanistic codon models with the variation of selective constraints attain slightly lower ΔAIC and ΔBIC than the corresponding models with rate variation, although the differences of ΔBIC between the two models are insignificant in the separating analyses of the genes. The phylogenetic trees of the datasets human-mtDNA and nDNA consist of extremely short branches only, in which nonsynonymous substitutions insignificantly occur under strong selective constraints. In such a phylogenetic tree, it is hard to estimate correctly the variation of selective constraint over sites as indicated by the high performance of the Equal-Constraint model. This would be the reason why the differences of ΔBIC between the mechanistic codon models of the dG4r and the dG4s are insignificant in the separating analyses of genes for the human-mtDNA. On the other hand, the present result for the nDNA, which consists of 10 genes that are not necessarily closely-located in the same chromosome, may indicate the possibility of rate variation over sites.

Site dependencies of selective constraints

Selective constraints against amino acid replacements at each site must reflect both structural and functional constraints on a

Table 6. Comparisons between various codon substitution models in the concatenating analysis of nDNA.

Codon substitution model[a]	**K**[b]	$\Delta\ell^c$	$\Delta\mathrm{AIC}^c$	$\Delta\mathrm{BIC}^c$	$\langle e^{w_{ab}}\rangle^{de}$	$\hat{\sigma}^{ef}$	$\hat{m}^{eg}$	[h]	$\hat{\alpha}^i$
LG-1-F-dG4r	62	−56.9	111.7	106.7		(0.0)			0.429
WAG-1-F-dG4r	62	−55.8	109.6	104.6		(0.0)			0.431
JTT-1-F-dG4r	62	−42.7	83.3	78.3		(0.0)			0.434
LG-1-F-dG4s	62	−82.5	163.1	158.1		(0.0)			0.102
WAG-1-F-dG4s	62	−81.0	160.1	155.0		(0.0)			0.103
JTT-1-F-dG4s	62	−68.6	135.3	130.3		(0.0)			0.110
KHG-2-F-dG4r	63	20.5	−41.1	−41.1		0.082			0.472
KHG-2-F-dG4s	63	0.0	0.0	0.0		0.214			0.118
No-Constraint-3-F-dG4r	64	−92.9	187.9	192.9	(1.0)	0.000	0.000	4.368	0.499
Equal-Constraint-4-F-dG4r	65	137.1	−270.3	−260.2	0.083	0.000	0.020	2.638	0.465
KHG-ML200-5-F-dG4r	66	143.2	−280.4	−265.4	0.085	0.000	0.018	2.500	0.461
EI-5-F-dG4r	66	144.6	−283.2	−268.1	0.061	0.000	0.026	2.675	0.458
LG-ML91+-5-F-dG4r	66	146.6	−287.1	−272.1	0.065	0.000	0.025	2.756	0.456
JTT-ML91+-5-F-dG4r	66	147.3	−288.5	−273.5	0.050	0.000	0.027	2.515	0.463
WAG-ML91+-5-F-dG4r	66	147.7	−289.3	−274.3	0.053	0.000	0.025	2.737	0.459
Equal-Constraint-4-F-dG4s	65	119.4	−234.9	−224.9	0.092	0.003	0.032	2.470	0.109
KHG-ML200-5-F-dG4s	66	124.4	−242.7	−227.7	0.080	0.000	0.039	2.333	0.157
EI-5-F-dG4s	66	125.9	−245.7	−230.7	0.060	0.000	0.049	2.450	0.152
LG-ML91+-5-F-dG4s	66	125.9	−245.7	−230.7	0.063	0.000	0.050	2.466	0.172
JTT-ML91+-5-F-dG4s	66	128.0	−250.1	−235.0	0.051	0.000	0.053	2.344	0.133
WAG-ML91+-5-F-dG4s	66	128.4	−250.7	−235.7	0.056	0.000	0.049	2.503	0.127
WAG-ML91+-4s-F-dG4r	65	146.1	−288.3	−278.2	0.055	0.016	(0.0)	2.755	0.449
WAG-ML91+-4-F-dG4r	65	147.7	−291.3	−281.3	0.053	(0.0)	0.025	2.737	0.459
WAG-ML91+-4s-F-dG4s	65	127.5	−250.9	−240.9	0.057	0.079	(0.0)	2.572	0.133
WAG-ML91+-4-F-dG4s	65	128.4	−252.8	−242.7	0.056	(0.0)	0.049	2.507	0.129
WAG-ML91+-5-F	65	109.7	−215.4	−205.4	0.055	0.001	0.059	2.535	

[a] In the models specified with the suffix "-3-", "-4-", "-4s-" or "-5-", three, four or five parameters are optimized with $f^{mut_\xi}=0.25$ and $m_{ag}/m_{tc|ag}=m_{ta}/m_{[tc][ag]}=m_{tg}/m_{[tc][ag]}=m_{ca}/m_{[tc][ag]}=1.0$. The prior probability of each category for the "dG4s" is $p(\Gamma_1)=0.75$, $p(\Gamma_2)=0.125$, and $p(\Gamma_3)=p(\Gamma_4)=0.0625$.
[b] The number of adjustable parameters.
[c] Differences from the reference state; $\Delta\ell=\ell+6739.3$, $\Delta\mathrm{AIC}=(-2\ell+2K)-13604.5$, and $\Delta\mathrm{BIC}=(-2\ell+K\log 1112)-13920.4$.
[d] The average of e^{w}_{ab} over all amino acid pairs $\{a,b\}$; $\langle e^{w_{ab}}\rangle\equiv\frac{1}{190}\sum_a\sum_{b>a}e^{w_{ab}}$.
[e] The value parenthesized means that the parameter is fixed at the value specified.
[f] The scale parameter of a Γ distribution for the variation of mutation rate over time.
[g] The ratio of double to single and of triple to double nucleotide change exchangeability; $\hat{m}\equiv\hat{m}_{[tc][ag]}$.
[h] The ratio of mean transitional to mean transversional exchangeability; $\hat{m}_{tc|ag}/\hat{m}_{[tc][ag]}$.
[i] The shape parameter of a discrete gamma distribution for the variation of mutation rate or selective constraint over sites.
doi:10.1371/journal.pone.0028892.t006

residue type at each site, which are required for a protein to fold into a unique native structure and to properly function, and vary among residue sites in a protein. Here a simple analysis of site dependencies of selective constraints has been performed to ascertain the correlation between selective constraints and structural constrains at each site.

Site dependencies of selective constraints are evaluated [40] as a posterior mean of $\langle e^{w_{i,ab}}\rangle(\equiv\sum_a\sum_{b>a}e^{w_{i,ab}}/190)$ over categories i for each site. Residue sites are categorized by the number of van der Waals contacts with surrounding non-solvent atoms in a protein structure, which are supposed to reflect the strength of structural constraints; neighboring residues along a polypeptide chain are not counted. Then, the posterior mean of $\langle e^{w_{i,ab}}\rangle$ are averaged over sites in each residue category and its dependence on the category is examined. In Fig. 3, the site dependencies of selective constrains are shown for the photosystem II CP47 chloroplast protein (psbB gene) and for the cytochrome c oxidase subunit 1 mitochondrial protein (COX1 gene). The van der Waals contacts were evaluated for the psbB in the 38-meric state of the photosystem II protein complex and for the COX1 in the biological 26-meric state of bovine heart cytochrome C oxidase in the fully reduced state; the protein coordinates 3ARC and 2EIJ in the PDB database were used. The posterior mean of selective constrains for each site was calculated in the LG-ML91+-12-F-dG4s for the concatenated sequences of the datasets cpDNA-9 and mammalian-mtDNA. It is clear that the selective constraints tend to be stronger at residues surrounded by more atoms, indicating that they reflect structural constraints at each residue site in a protein. Here we have taken account of purifying selection only, but positive selection can be also examined [29] in terms of $\exp w_{ab}$ (fixation rate) at each site.

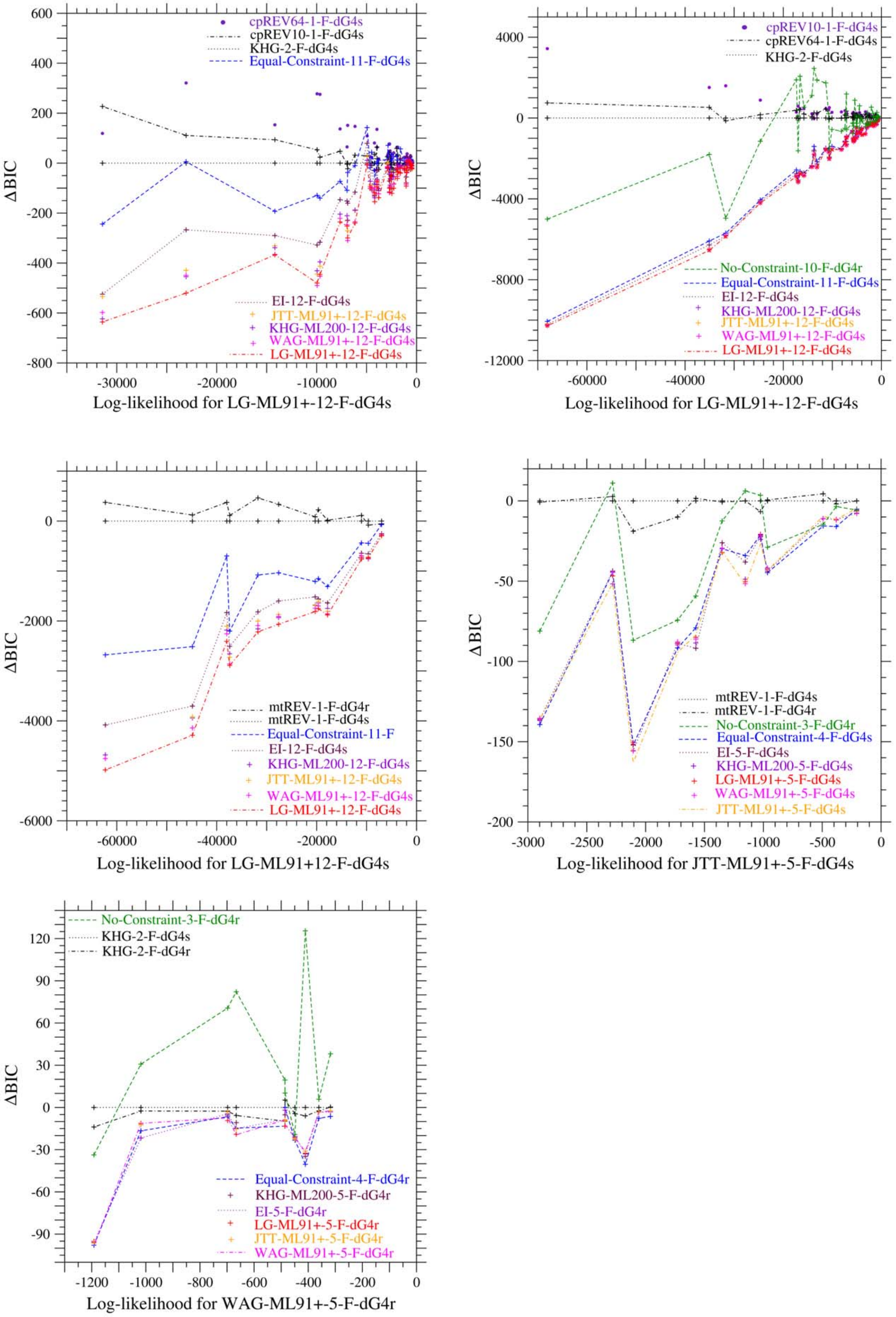
ΔBIC
Log-likelihood for LG-ML91+-12-F-dG4s
600
400
200
0
-200
-400
-600
-800
-30000
-20000
-10000
0
cpREV64-1-F-dG4s
cpREV10-1-F-dG4s
KHG-2-F-dG4s
Equal-Constraint-11-F-dG4s
EI-12-F-dG4s
JTT-ML91+-12-F-dG4s
KHG-ML200-12-F-dG4s
WAG-ML91+-12-F-dG4s
LG-ML91+-12-F-dG4s
ΔBIC
Log-likelihood for LG-ML91+-12-F-dG4s
4000
0
-4000
-8000
-12000
-60000
-40000
-20000
0
cpREV10-1-F-dG4s
cpREV64-1-F-dG4s
KHG-2-F-dG4s
No-Constraint-10-F-dG4r
Equal-Constraint-11-F-dG4s
EI-12-F-dG4s
KHG-ML200-12-F-dG4s
JTT-ML91+-12-F-dG4s
WAG-ML91+-12-F-dG4s
LG-ML91+-12-F-dG4s
ΔBIC
Log-likelihood for LG-ML91+12-F-dG4s
0
-2000
-4000
-6000
-60000
-40000
-20000
0
mtREV-1-F-dG4r
mtREV-1-F-dG4s
Equal-Constraint-11-F
EI-12-F-dG4s
KHG-ML200-12-F-dG4s
JTT-ML91+-12-F-dG4s
WAG-ML91+-12-F-dG4s
LG-ML91+-12-F-dG4s
ΔBIC
Log-likelihood for JTT-ML91+-5-F-dG4s
0
-50
-100
-150
-200
-3000
-2500
-2000
-1500
-1000
-500
0
mtREV-1-F-dG4s
mtREV-1-F-dG4r
No-Constraint-3-F-dG4r
Equal-Constraint-4-F-dG4s
EI-5-F-dG4s
KHG-ML200-5-F-dG4s
LG-ML91+-5-F-dG4s
WAG-ML91+-5-F-dG4s
JTT-ML91+-5-F-dG4s
ΔBIC
Log-likelihood for WAG-ML91+-5-F-dG4r
120
90
60
30
0
-30
-60
-90
-1200
-1000
-800
-600
-400
-200
0
No-Constraint-3-F-dG4r
KHG-2-F-dG4s
KHG-2-F-dG4r
Equal-Constraint-4-F-dG4r
KHG-ML200-5-F-dG4r
EI-5-F-dG4r
LG-ML91+-5-F-dG4r
JTT-ML91+-5-F-dG4r
WAG-ML91+-5-F-dG4r

Figure 1. Comparisons of $\Delta \mathrm{BIC}$ of each gene in each dataset among models. $\Delta \mathrm{BIC}$ of each gene in cpDNA-9 (A), cpDNA-55 (B), mammalian-mtDNA (C), human-mtDNA (D), and nDNA (E) for each specified model is plotted against its log-likelihood value for the best model in the concatenating analysis of the genes. The horizontal dotted line of $\Delta \mathrm{BIC} = 0$ shows the reference model for each dataset. The best model is shown by the lowest dot-dashed line. The lower dotted line in each figure shows the data points for the EI model. The lower and the upper broken lines show the Equal-Constraint and the No-Constraint models, respectively. The No-Constraint model is not shown for cpDNA-9 and mammalian-mtDNA, because its $\Delta \mathrm{BIC}$ values are too large. In the models specified with the suffix "-5-" for human-mtDNA and nDNA, five parameters were optimized with $f^{\mathrm{mut}_\xi} = 0.25$ and $m_{ag}/m_{tc|ag} = m_{ta}/m_{[tc][ag]} = m_{tg}/m_{[tc][ag]} = m_{ca}/m_{[tc][ag]} = 1.0$.
doi:10.1371/journal.pone.0028892.g001

Estimates of branch lengths under different models

The correct estimation of branch length is critical for the estimations of phylogeny and divergence times. It is known that branch-length estimation is significantly influenced by model selection. Yang et al. [33] found that branch lengths are severely underestimated by nucleotide substitution models in which rate variation over site is ignored. Also they found that simpler and worse models tend to underestimate branch lengths more severely, and such a bias is more serious for longer branches.

In Fig. 4, branch lengths estimated by the models with a uniform substitution rate, with the variation of selective constraint, and with the variation of mutation rate over sites are plotted against those estimated by the model with the variation of selective constraint over sites. The dotted lines in these figures are ones connecting the origin and the point of the longest branch on the abscissa. Assuming the variation of mutation rate or selective constraint leads to longer estimates of branch lengths than the uniform substitution rate over sites. However, the estimates of branch lengths are significantly different between the schemes of variable mutation rates and of variable selective constraints over sites, and assuming the variation of mutation rate estimates branch lengths much longer for all datasets than the variation of selective constraint.

Branch lengths estimated by the models with the variation of mutation rate (dG4r) and with a uniform substitution rate are both roughly proportional to those shown on the abscissa, i.e., those estimated by the model with the variation of selective constraint (dG4s). However, as pointed out by Yang et al. [33], a systematic bias in the estimation of branch length is shown; the ratio of the branch length estimate of a worse model to that of the best model tends to be smaller for longer branches irrespective of overestimation or underestimation. For cpDNA-9, cpDNA-55 and mammalian-mtDNA, for which the dG4s is the best model, plus marks for a uniform substitution rate and cross marks for the dG4r are plotted in a concave pattern, although the concave pattern for dG4r is not clear in cpDNA-55. For nDNA, which the dG4r fits better than the dG4s, cross marks for the dG4r are plotted in a slightly convex pattern. This systematic bias indicates that the worse models tend to underestimate the frequencies of multiple substitutions in long branches in comparison with short branches.

When the different types of models are compared with each other, the correlation of branch lengths between the models is not always good. In Fig. 5, the estimates of branch lengths for cpDNA-9 and mammalian-mtDNA in the Equal-constraint model and in the amino acid substitution model converted into the codon models are plotted against those in the best model. These estimates for cpDNA-9 are roughly proportional to those in the best model, although there is a systematic bias. However, the correlation of branch lengths between the mtREV-1-F-dG4s and the best model for mammalian-mtDNA is not as good as those between the models for cpDNA-9.

In the result, except for the datasets consisting of highly-homologous sequences, the variation of selective constraint is a better model than the variation of mutation rate, and assuming the variation of mutation rate leads to the overestimation of branch length. Even for highly-homologous sequence families, the model with the variation of selective constraint may not be too bad, because the differences of AIC and BIC between the models with variable mutation rates and with variable selective constraints are not significantly large, and the branch lengths estimated by those models are almost proportional to each other.

Multiple nucleotide changes in infinitesimal time

Codon substitutions requiring multiple nucleotide changes can be caused by either multiple steps of single nucleotide changes or single steps of multiple nucleotide changes. In the present mechanistic codon substitution model, codon mutations by multiple nucleotide changes in infinitesimal time are taken into account. The mechanistic codon substitution models with the various selective constraint matrices all indicate $m(\equiv m_[tc][ag]) > 0$ for the datasets cpDNA-9, cpDNA-55, and mammalian-mtDNA, which include long branches. The $\Delta \mathrm{AIC}$ and $\Delta \mathrm{BIC}$ values consistently indicate that the model LG-ML91+-12-F-dG4s, in which multiple nucleotide changes are allowed, fits these datasets better than the model LG-ML91+-11s-F-dG4s, in which multiple nucleotide changes are disallowed. Also, the LRTs for LG-ML91+-11s-F-dG4s nested by LG-ML91+-12-F-dG4s reject the assumption of single nucleotide changes with $\mathrm{p-value} \ll 0.00001$ for these datasets; see Tables 2, 3, 4. This result is consistent with a report [19] that the mechanistic codon model could not well fit observed substitution frequency data unless multiple nucleotide changes in infinitesimal time are allowed.

On the other hand, the parameter for multiple nucleotide changes is not significant for the datasets human-mtDNA and nDNA that consist of closely-related or highly-conserved sequences, and whose phylogenetic trees consist of short branches only. This fact indicates that multiple nucleotide changes rarely occur in short evolutionary periods, and multiple nucleotide changes detected in relatively long branches of cpDNA-9, cpDNA-55, and mammalian-mtDNA may result from compensatory substitutions that shortly succeed single nucleotide substitutions, or other mechanisms. A possibility of successive single compensatory substitutions for multiple nucleotide changes was pointed out by Bazykin et al. [26]. Whatever results in multiple nucleotide changes in long evolutionary periods, the present method, in which multiple nucleotide changes in infinitesimal time are allowed, for codon substitutions is effective to improve the likelihood of a phylogenetic tree with long branches.

Variation of mutation rate over time

The site-specific variation of amino acid substitution rate over time was first discussed as a covarion model by Fitch and Markowitz [35], and recently its significance have been indicated again for rRNA [36] and cytochrome *b* [37]. Although amino acid substitutions may occur in a concerted manner with other interacting sites, causing the variation of selective constraint over time, here we has examined the variation of mutation rate over time at each site.

The present model includes a parameter σ for the variation of mutation rate over time. The scale factor $\sigma = 0$ for a Γ distribution

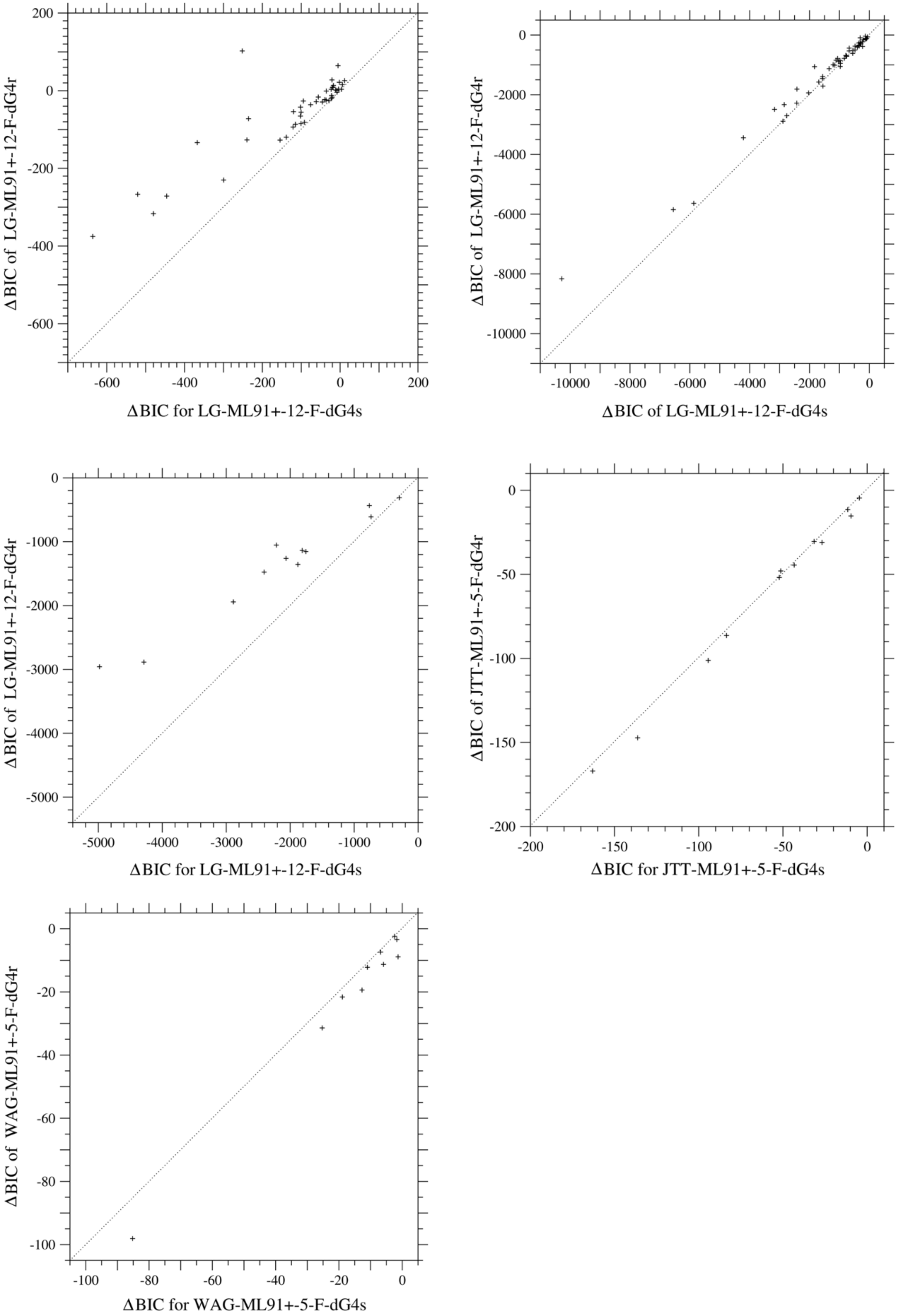
200
0
-200
-400
-600
ΔBIC of LG-ML91+-12-F-dG4r
-600
-400
-200
0
200
ΔBIC for LG-ML91+-12-F-dG4s
0
-2000
-4000
-6000
-8000
-10000
ΔBIC of LG-ML91+-12-F-dG4r
-10000
-8000
-6000
-4000
-2000
0
ΔBIC of LG-ML91+-12-F-dG4s
0
-1000
-2000
-3000
-4000
-5000
ΔBIC of LG-ML91+-12-F-dG4r
-5000
-4000
-3000
-2000
-1000
0
ΔBIC for LG-ML91+-12-F-dG4s
0
-50
-100
-150
-200
ΔBIC of JTT-ML91+-5-F-dG4r
-200
-150
-100
-50
0
ΔBIC for JTT-ML91+-5-F-dG4s
0
-20
-40
-60
-80
-100
ΔBIC of WAG-ML91+-5-F-dG4r
-100
-80
-60
-40
-20
0
ΔBIC for WAG-ML91+-5-F-dG4s

Figure 2. Comparison of $\Delta\mathrm{BIC}$ of each gene in each dataset between the models with variable mutation rates and with variable selective constraints over sites. $\Delta\mathrm{BIC}$ of each gene in cpDNA-9 (A), cpDNA-55 (B), mammalian-mtDNA (C), human-mtDNA (D), and nDNA (E) is compared between the models with the variation of mutation rate (dG4r) and with the variation of selection constraint (dG4s) over sites. The dotted line shows the line of equal values between the ordinate and the abscissa. In the models specified with the suffix "-5-" for human-mtDNA, five parameters were optimized with $f^{\mathrm{mut}_\xi}=0.25$ and $m_{ag}/m_{tc|ag}=m_{ta}/m_{[tc][ag]}=m_{tg}/m_{[tc][ag]}=m_{ca}/m_{[tc][ag]}=1.0$.
doi:10.1371/journal.pone.0028892.g002

means no variation of mutation rate over time. Because the present simple approximation works by replacing the substitution matrix S by its expected value $E(S)$ under rate variations, the parameter σ will not only reflect the variation of mutation rate over time but also be affected by the variations of selective constraints over time and of substitution rate over sites, especially if both the variations of mutation rate and of selective constraint over sites are not taken into account; it tends to take larger values in models assuming a uniform rate over sites than variable mutation rates or selective constraints. Also, if only single nucleotide changes in infinitesimal time are assumed, i.e., $m(\equiv m_{[tc][ag]})=0$, this parameter (σ) will be estimated to be larger to increase the probability of multiple steps of substitutions. The reverse is also true.

The mechanistic codon substitution models specified with a suffix dG4s, in which selective constraints are variable across sites, all indicate $\hat{\sigma}>0$ for the datasets cpDNA-9, cpDNA-55, and mammalian-mtDNA, which include long branches. The $\Delta\mathrm{AIC}$ and $\Delta\mathrm{BIC}$ values indicate that the model LG-ML91+-12-F-dG4s including σ as a parameter fits these datasets better than the model LG-ML91+-11-F-dG4s assuming $\sigma=0$. Also, the LRTs for LG-ML91+-11-F-dG4s nested by LG-ML91+-12-F-dG4s reject a constant mutation rate over time with $\mathrm{p-value} \ll 0.00001$ for all cpDNA-9, cpDNA-55, and mammalian-mtDNA; see Tables 2, 3, 4. Therefore, rate variation over time should not be ignored for highly-diverged sequences. The ML estimate of σ for mammalian-mtDNA is larger than 1, while it is less than 0.5 for the other two datasets. The variation of mutation rate among lineages in primate mtDNAs has been indicated [6,55].

As shown in Tables 2, 3, 4, 5, 6, when mutation rates are assumed to be variable across sites, i.e., in the mechanistic codon substitution models specified with a suffix dG4r, the parameter σ has been estimated to be almost equal to zero for all the datasets, even for the datasets cpDNA-9, cpDNA-55, and mammalian-mtDNA, for which the models assuming variable selective constraints indicate $\hat{\sigma}>0$. Variable mutation rates across sites are taken into account in such a way that each site has multiple mutation rates with certain probabilities given by a discrete gamma distribution. Thus, in the present approximation it would be hard to distinguish the variation of mutation rate over time at each site from that over sites in these models.

Transition/transversion bias

One of the advantages in mechanistic codon substitution models over amino acid substitution models is that mutational tendencies at the nucleotide level can be estimated. The estimation of mutational tendencies by mechanistic codon substitution models must be more precise than by nucleotide substitution models

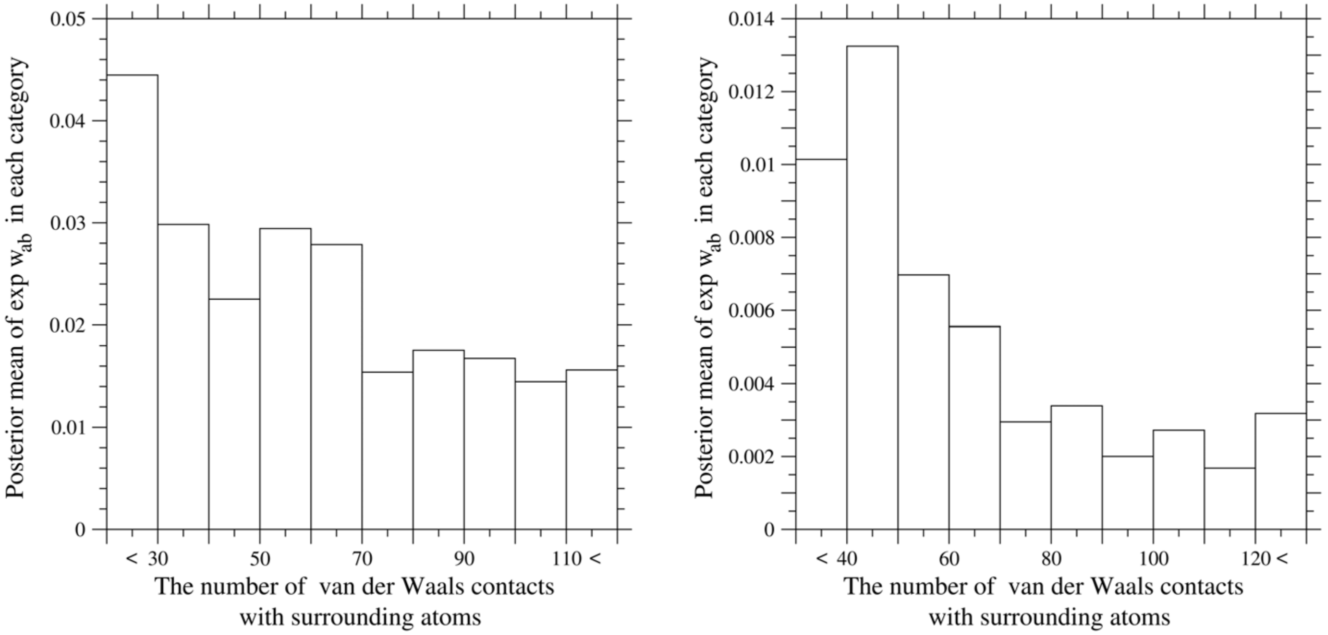

Figure 3. Site dependences of selective constraints. Site dependences of selective constraints in the photosystem II CP47 protein (psbB gene) (A) and cytochrome c oxidase subunit 1 mitochondrial protein (COX1 gene) (B) are shown. Residue sites are categorized by the number of van der Waals contacts with surrounding non-solvent atoms in the protein structure; neighboring residues along a polypeptide chain are not counted. The degree of van der Waals contact for an atom pair, which is separated by r and whose van der Waals distance is equal to r_m, is defined as $2(r_m/r)^6-(r_m/r)^{12}$ for $r_m/r<1$ and 1 for $r_m/r\geq 1$. The van der Waals contacts are evaluated for the psbB in the 38-meric state of the photosystem II complex from *Thermosynechococcus vulcanus*, and for the COX1 in the biological 26-meric state of bovine heart cytochrome C oxidase in the fully reduced state; the protein coordinates 3ARC and 2EIJ in the PDB database were used. Posterior mean of selective constrains ($\langle e^{w_{i,ab}} \rangle$) averaged over sites in each residue category is shown in the ordinate. The posterior mean of selective constrains were calculated by the LG-ML91+-12-F-dG4s for the concatenated sequences of the datasets cpDNA-9 and mammalian-mtDNA.
doi:10.1371/journal.pone.0028892.g003

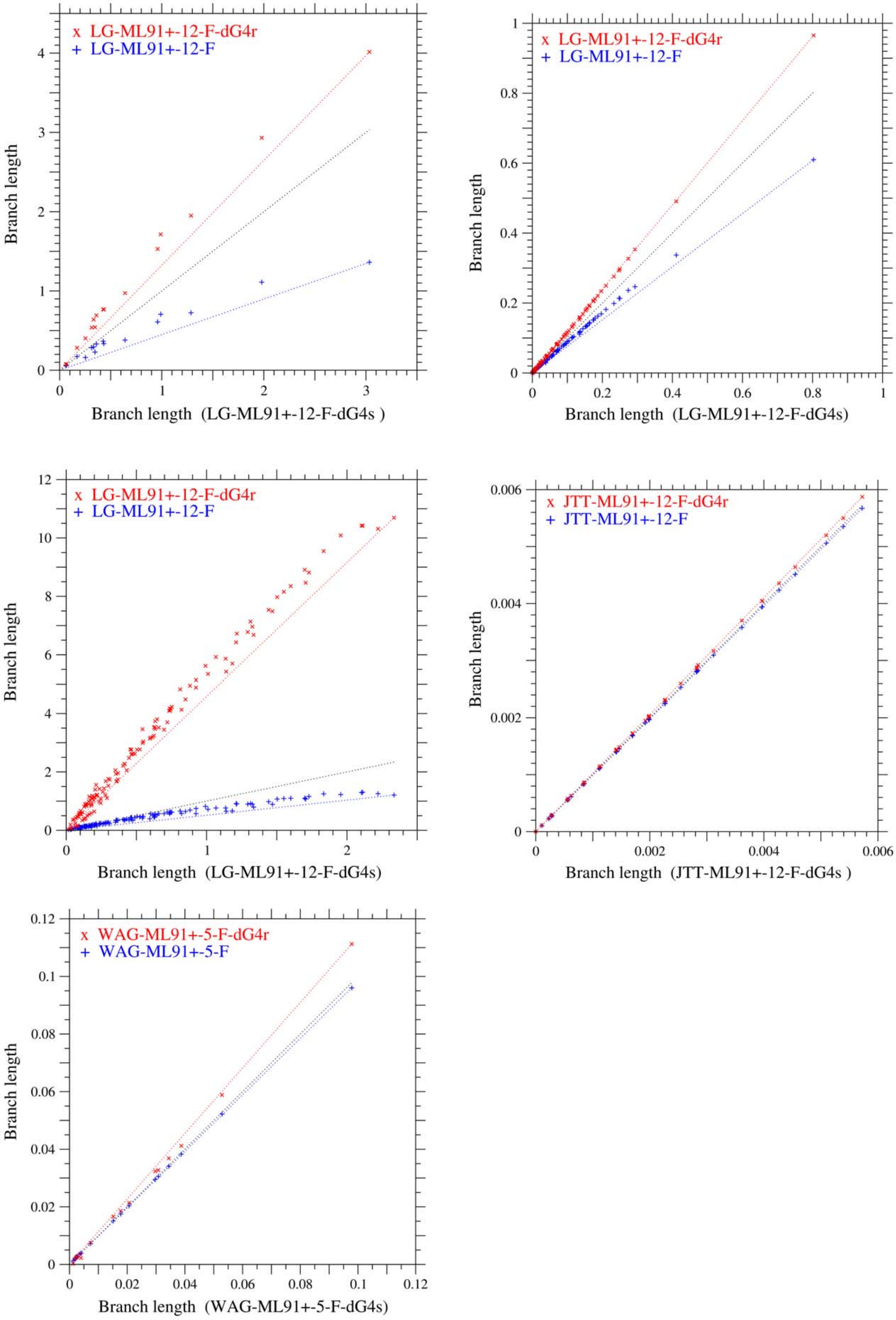
x LG-ML91+-12-F-dG4r
+ LG-ML91+-12-F
Branch length
Branch length (LG-ML91+-12-F-dG4s)
0
1
2
3
4
x LG-ML91+-12-F-dG4r
+ LG-ML91+-12-F
Branch length
Branch length (LG-ML91+-12-F-dG4s)
0
0.2
0.4
0.6
0.8
1
x LG-ML91+-12-F-dG4r
+ LG-ML91+-12-F
Branch length
Branch length (LG-ML91+-12-F-dG4s)
0
2
4
6
8
10
12
x JTT-ML91+-12-F-dG4r
+ JTT-ML91+-12-F
Branch length
Branch length (JTT-ML91+-12-F-dG4s)
0
0.002
0.004
0.006
x WAG-ML91+-5-F-dG4r
+ WAG-ML91+-5-F
Branch length
Branch length (WAG-ML91+-5-F-dG4s)
0
0.02
0.04
0.06
0.08
0.1
0.12

Figure 4. Comparisons of branch lengths estimated by the models with a uniform rate, variable mutation rates, and variable selective constraints over sites. Branch lengths estimated for the phylogenetic trees of cpDNA-9 (A), cpDNA-55 (B), mammalian-mtDNA (C), human-mtDNA (D), and nDNA (E) are compared among models. The abscissa shows the branch lengths estimated by the model with the variation of selection constraint (dG4s). The LG-ML91+-12-F-dG4s is the best model except for human-mtDNA and nDNA. The best model is JTT-ML91+-12-F-dG4s for human-mtDNA and WAG-ML91+-5-F-dG4r for nDNA. The models with the variation of mutation rate (dG4r) and with a uniform substitution rate over sites are shown by cross and plus marks, respectively. The model with the variation of selection constraint (dG4s) is shown by the middle dotted line. The dotted lines in each figure are ones connecting the origin and the respective estimates for the longest branch on the abscissa. In the models specified with the suffix ''-5-'' for human-mtDNA and nDNA, five parameters were optimized with $f^{\mathrm{mut}_\xi}=0.25$ and $m_{ag}/m_{tc|ag}=m_{ta}/m_{[tc][ag]}=m_{tg}/m_{[tc][ag]}=m_{ca}/m_{[tc][ag]}=1.0$.
doi:10.1371/journal.pone.0028892.g004

applied to all codon positions, because selection at the amino acid level is taken into account.

Transitional substitutions have been noted to occur more frequently than transversions [56,57], and transition/transversion rate bias is more pronounced in animal mitochondrial DNAs than in nuclear or chloroplast DNAs [58]. Different measures have been used for transition to transversion bias [55,58,59]. One is the ratio of transitional differences to transversional differences between two sequences. Another is the ratio of the total transitional to the total transversional rate. Also, the ratio of transitional to transversional substitution exchangeability has been used. Here, the ratio of the mean transitional to the mean transversional exchangeability is used, because each type of transitional and transversional mutations occurs with a different exchangeability. The ratio ($\hat{m}_{tc|ag}/\hat{m}_{[tc][ag]}$) of the mean transitional to the mean transversional exchangeability is listed in Tables 2, 3, 4, 5, 6 for all datasets. The values of $\hat{m}_{tc|ag}/\hat{m}_{[tc][ag]}$ in the mechanistic codon substitution models with the various estimates of selective constraints fall into a narrow range for each dataset. They range from 3.7 to 7.2 for mammal-mtDNA, and from 30.5 to 39.0 for human-mtDNA. On the other hand, they fall into the range of much smaller values from 1.7 to 2.4 for cpDNA-9, from 2.4 to 2.7 for cpDNA-55, and from 2.3 to 2.8 for nDNA. The ratio of the mean transitional to the mean transversional exchangeability is estimated to be almost 10–20 times larger for human mitochondrial DNA but only 2–3 times larger for mammalian mitochondrial DNA than for nuclear and chloroplast DNAs. Adachi and Hasegawa [6] reported that the transitional mutation rate and the ratio of transitional to transversional mutation rate at four-fold degenerate sites of mtDNA were higher by about two times in humans than in apes. On the other hand, Yang and Yoder [55] showed that the maximum likelihood estimate of the ratio of transitional to transversional substitution rate changes with the species included in the analysis, and was always larger at low than at high sequence divergence. It was suggested [55] that the variable rates of transitional and transversional mutations among evolutionary lineages might cause such a sample dependence.

Conclusions

In the present mechanistic codon substitution model, single nucleotide mutations are modeled by the GTR model and multiple nucleotide mutations in infinitesimal time are assumed to occur independently at each position of codon, and selective constraints on amino acids are approximated by a linear function of the empirical selective constraints. It has been shown that even the Equal-Constraint model performs far better for a wide range of sequences from highly-homologous to highly-diverged sequences than both the No-Constraint model and the amino acid

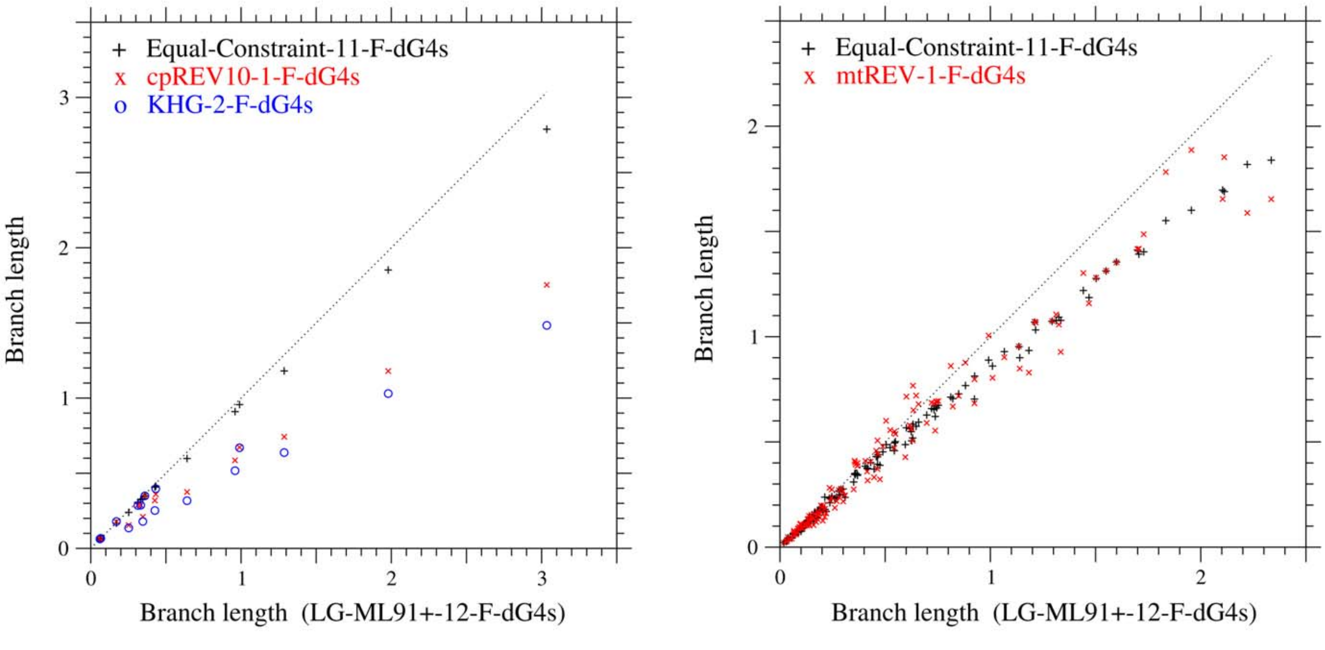

Figure 5. The estimates of branch lengths for the phylogenetic tree of each dataset under the different types of models. Branch lengths estimated for cpDNA-9 (A) and mammalian-mtDNA (B) are compared among models. The abscissa shows the branch lengths estimated by the best model with the variation of selective constraint, LG-ML91+-12-F-dG4s. The dotted line in each figure shows branch lengths estimated by the best model shown on the abscissa.
doi:10.1371/journal.pone.0028892.g005

substitution models converted into the codon substitution models. The No-Constraint model is a nucleotide substitution model extended to allow multiple nucleotide changes in infinitesimal time. On the other hand, the codon substitution model converted from the amino acid substitution model is extended here in such a way that the special case of $w_0 = -\infty$ is exactly equivalent to the amino acid substitution model [17]. Thus, the performance of the Equal-Constraint model indicates that codon substitution models are superior to nucleotide and amino acid substitution models.

The present analyses have also shown that the mechanistic models with the amino acid dependent selective constraints do not only perform far better especially for phylogenetic trees consisting of relatively long branches than the Equal-Constraint model, but better even for phylogenetic trees consisting of short branches. This result indicates the superiority of the selective constraint matrices (w_{ab}^{estimate}) estimated by maximizing the respective likelihoods of the observed substitution frequency matrices of 1-PAM [19]. In long branches, nonsynonymous substitutions increase, and therefore the proper evaluation of selective constraints on amino acids becomes critical. On the other hand, in short branches in which nonsynonymous substitutions are insignificant, the proper evaluation of mutational tendencies at the nucleotide level becomes important. The former is the situation in which amino acid substitution models perform better than nucleotide substitution models. Inversely, the latter is the situation in which nucleotide substitution models perform better, although they are not superior for slow-evolving proteins, because there is a possibility that synonymous substitutions are saturated even in short branches; the dataset nDNA is an example of such a case. However, mutational tendencies at the nucleotide level and the strength of selective constraints cannot be tailored to each gene in the amino acid substitution models, and selection on amino acid replacements cannot be taken into account in the nucleotide substitution models. Thus, mechanistic codon models that can tailor both mutational tendencies and the strength of selective constraints are superior to both nucleotide and amino acid substitution models.

It was pointed out [18] that codon substitution models require intensive computation to recalculate eigenvalues and eigenvectors of a 64-dimensional matrix. Simultaneous optimizations of a tree topology and model parameters may be hard. However, model parameters may be fixed at the values estimated for one of the reasonable trees, because the optimum values of model parameters do not severely depend on a tree topology, unless tree topologies are unrealistic. On the other hand, the mechanistic codon substitution model can provide much information on mutational tendencies and the strength of selective constraints. In addition, the present model enables us to distinguish the variations of mutation rate and of selective constraint over sites. The variation of mutation rate over time can also be discussed.

The present analyses show that multiple nucleotide changes in infinitesimal time are statistically significant in long branches as well as the variation of mutation rate over time. It has been also shown that the variation of amino acid substitution rate over sites results from variable selective constraints rather than variable mutation rates at least in the phylogenetic trees of cpDNA-9, cpDNA-55, and mammalian-mtDNA including long branches. Branch lengths will be overestimated for these datasets if the variation of mutation rate over sites is assumed instead of the variation of selective constraint. The capability of the mechanistic codon substitution models to extract biological knowledge from protein-coding sequences makes them superior to both nucleotide and amino acid substitution models.

Acknowledgments

We would like to thank Dr. Hidenori Nishihara and Dr. Tae-Kun Seo for kindly providing sequence data. We also thank Dr. Bojian Zhong and Prof. Masami Hasegawa for providing the amino acid substitution probability matrix of cpREV64.

Author Contributions

Conceived and designed the experiments: SM. Performed the experiments: SM. Analyzed the data: SM. Contributed reagents/materials/analysis tools: SM. Wrote the paper: SM.

References

1. Kimura M (1980) A simple model for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16: 111–120.
2. Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22: 160–174.
3. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512–526.
4. Dayhoff MO, Schwartz RM, Orcutt BC (1978) A model of evolutionary change in proteins. In: Dayhoff MO, ed. Atlas of protein sequence and structure. Washington D.C.: National Biomedical Research Foundation, volume 5. Suppl. 3 edition. pp 345–352.
5. Jones DT, Taylor WR, Thornton JM (1992) The rapid generation of mutation data matrices from protein sequences. CABIOS 8: 275–282.
6. Adachi J, Hasegawa M (1996) Model of amino acid substitution in proteins encoded by mitochondrial DNA. J Mol Evol 42: 459–468.
7. Yang Z, Nielsen R, Hasegawa M (1998) Models of amino acid substitution and application to mitochondrial protein evolution. Mol Biol Evol 15: 1600–1611.
8. Adachi J, Waddell PJ, Martin W, Hasegawa M (2000) Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol 50: 348–358.
9. Dimmic MW, Mindell DP, Goldstein RA (2000) Modelling evolution at the protein level using an adjustable amino acid fitness model. Pacific Symposium on Biocomputing 5: 18–29.
10. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol 18: 691–699.
11. Le SQ, Gascuel O (2008) An improved general amino acid replacement matrix. Mol Biol Evol 25: 1307–1320.
12. Huelsenbeck JP, Joyce P, Lakner C, Ronquist F (2008) Bayesian analysis of amino acid substitution models. Phil Trans R Soc B 363: 3941–3953.
13. Shapiro B, Rambaut A, Drummond AJ (2006) Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol Biol Evol 23: 7–9.
14. Schneider A, Cannarozzi GM, Gonnet GH (2005) Empirical codon substitution matrix. BMC Bioinformatics 6: 134.
15. Kosiol C, Holmes I, Goldman N (2007) An empirical codon model for protein sequence evolution. Mol Biol Evol 24: 1464–1479.
16. Delport W, Scheffler K, Gravenor MB, Muse SV, Kosakovsky Pond S (2010) Benchmarking multirate codon models. PLoS One 5: e11587.
17. Seo TK, Kishino H (2008) Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst Biol 57: 367–377.
18. Seo TK, Kishino H (2009) Statistical comparison of nucleotide, amino acid, and codon substitution models for evolutionary analysis of protein-coding sequences. Syst Biol 58: 199–210.
19. Miyazawa S (2011) Selective constraints on amino acids estimated by a mechanistic codon substitution model with multiple nucleotide changes. PLoS One 10: 1371.
20. Whelan S, Goldman N (2004) Estimating the frequency of events that cause multiple-nucleotide changes. Genetics 167: 2027–2043.
21. Doron-Faigenboim A, Pupko T (2007) A combined empirical and mechanistic codon model. Mol Biol Evol 24: 388–397.
22. Miyazawa S, Jernigan RL (1993) A new substitution matrix for protein sequence searches based on contact frequencies in protein structures. Protein Eng 6: 267–278.
23. Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA. Mol Biol Evol 11: 725–736.
24. Muse SV, Gaut BS (1994) Nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715–724.
25. Delport W, Scheffler K, Botha G, Gravenor MB, Muse SV, et al. (2010) Codontest: Modeling amino acid substitution preferences in coding sequences. PLoS Comp Biol 6: e1000885.

26. Bazykin G, Kondrashov F, Ogurtsov A, Sunyaev S, Kondrashov A (2004) Positive selection at sites of multiple amino acid replacements since rat-mouse divergence. Nature 429: 558–562.
27. Averof M, Rokas A, Wolfe KH, Sharp PM (2000) Evidence for a high frequency of simultaneous double-nucleotide substitutions. Science 287: 1283–1286.
28. Yang Z, Nielsen R (2008) Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol 25: 568–579.
29. Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431–449.
30. Halpern AL, Bruno WJ (1998) Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol 15: 910–917.
31. Yang Z (1993) Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Mol Biol Evol 10: 1396–1401.
32. Yang Z (1994) Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol 39: 306–314.
33. Yang Z, Goldman N, Friday A (1994) Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol Biol Evol 11: 316–324.
34. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ, eds. Evolving genes and proteins. New York: Academic Press. pp 97–116.
35. Fitch WM, Markowitz E (1970) An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. Biochem Genet 4: 579–593.
36. Galtier N (2001) Maximum-likelihood phylogenetic analysis under a covarion-like model. Mol Biol Evol 18: 866–873.
37. Lopez P, Casane D, Philippe H (2002) Heterotachy, an important process of protein evolution. Mol Biol Evol 19: 1–7.
38. Zhong B, Yonezawa T, Zhong Y, Hasegawa M (2010) The position of gnetales among seed plants: overcoming pitfalls of chloroplast phylogenomics. Mol Biol Evol 10: 1093.
39. Miyata T, Miyazawa S, Yasunaga T (1979) Two type of amino acid substitutions in protein evolution. J Mol Evol 12: 219–236.
40. Yang Z (1995) A space-time process model for the evolution of DNA sequences. Genetics 139: 993–1005.
41. Jansen RK, Cai Z, Raubeson LA, Daniell H, dePamphilis CW, et al. (2007) Analysis of 81 genes from 64 plastid genomes resolves relationships in angiosperms and identifies genome-scale evolutionary patterns. Proc Natl Acad Sci USA 104: 19369–19374.
42. Nikaido M, Cao Y, Harada M, Okada N, Hasegawa M (2003) Mitochondrial phylogeny of hedgehogs and monophyly of eulipotyphla. Mol Phylogenet Evol 28: 276–284.
43. Ingman M, Kaessmann H, Paabo S, Gyllensten U (2000) Mitochondrial genome variation and the origin of modern humans. Nature 408: 708–713.
44. Nishihara H, Okada N, Hasegawa M (2007) Rooting the eutherian tree: the power and pitfalls of phylogenomics. Genome Biol 8: R191.1–R191.10.
45. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, et al. (2007) Clustalw and clustalx version 2.0. Bioinformatics 23: 2947–2948.
46. Guindon S, Gascuel O (2003) Simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol 52: 696–704.
47. Posada D, Crandall KA (2001) Selecting the best-fit model of nucleotide substitution. Syst Biol 50: 580–601.
48. Stuart A, Ord K (1996) Likelihood ratio tests and the general linear hypothesis. In: Kendall's advanced theory of statistics. London: Edward Arnold, volume 2. 5th edition.
49. Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Contr AC-19: 716–723.
50. Schwarz G (1974) Estimating the dimension of a model. Ann Stat 6: 461–464.
51. Minin V, Abdo Z, Joyce P, Sullivan J (2003) Performance-based selection of likelihood models for phylogeny estimation. Syst Biol 52: 674–683.
52. Abdo Z, Minin VN, Joyce P, Sullivan J (2005) Accounting for uncertainty in the tree topology has little effect on the decision-theoretic approach to model selection in phylogeny estimation. Mol Biol Evol 22: 691–703.
53. Suchard MA, Weiss RE, Sinsheimer JS (2001) Bayesian selection of continuous-time markov chain evolutionary models. Mol Biol Evol 18: 1001–1013.
54. Go M, Miyazawa S (1980) Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution. Int J Peptide Protein Res 15: 211–224.
55. Yang Z, Yoder AD (1999) Estimation of the transition / transversion rate bias and species sampling. J Mol Evol 48: 274–283.
56. Brown WW, Prager EM, Wang A, Wilson AC (1982) Mitochondrial DNA sequences of primates: Tempo and mode of evolution. J Mol Evol 18: 225–239.
57. Gojobori T, Li WH, Grau D (1982) Patterns of nucleotide substitution in pseudogenes and functional genes. J Mol Evol 18: 360–369.
58. Wakeley J (1996) The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. Trends in Ecology and Evolution 11: 158–163.
59. Adachi J, Hasegawa M (1996) Tempo and mode of synonymous substitutions in mitochondrial DNA of primates. Mol Biol Evol 13: 200–208.