Statistical Comparison of Nucleotide, Amino Acid, and Codon Substitution Models for Evolutionary Analysis of Protein-Coding Sequences

TAE-KUN SEO1,* and HIROHISA KISHINO2

1Professional Programme for Agricultural Bioinformatics;
2Laboratory of Biometrics and Bioinformatics, Graduate School of Agricultural and Life Sciences,
University of Tokyo, 1-1-1 Yayoi Bunkyo-Ku, Tokyo 113-8657, Japan;

*Correspondence to be sent to: Professional Programme for Agricultural Bioinformatics, Graduate School of Agricultural and Life Sciences,
University of Tokyo, 1-1-1 Yayoi Bunkyo-Ku, Tokyo 113-8657, Japan; E-mail: seo@ia.a.u-tokyo.ac.jp.

Abstract.—Statistical models for the evolution of molecular sequences play an important role in the study of evolutionary processes. For the evolutionary analysis of protein-coding sequences, 3 types of evolutionary models are available: 1) nucleotide, 2) amino acid, and 3) codon substitution models. Selecting appropriate models can greatly improve the estimation of phylogenies and divergence times and the detection of positive selection. Although much attention has been paid to the comparisons among the same types of models, relatively little attention has been paid to the comparisons among the different types of models. Additionally, because such models have different data structures, comparison of those models using conventional model selection criteria such as Akaike information criterion (AIC) or Bayesian information criterion (BIC) is not straightforward. Here, we suggest new procedures to convert models of the above-mentioned 3 types to 64-dimensional models with nucleotide triplet substitution. These conversion procedures render it possible to statistically compare the models of these 3 types by using AIC or BIC. By analyzing divergent and conserved interspecific mammalian sequences and intraspecific human population data, we show the superiority of the codon substitution models and discuss the advantages and disadvantages of the models of the 3 types. [AIC; amino acid model; BIC; codon model; likelihood ratio test; model comparison; nucleotide model.]

The statistical models for the evolution of molecular sequences are designed to approximate the simplified forms of the complex aspects of evolutionary processes. In these simplified forms, evolutionary processes are summarized using a certain number of parameters. This simplification enables us to easily understand evolutionary processes and identify the major driving forces behind them. For the evolutionary analysis of nucleotide sequences, 3 types of evolutionary models are used: 1) nucleotide, 2) amino acid, and 3) codon substitution models. Depending on the type of data, all or some of these models are applicable. This study focuses on the analysis of protein-coding DNA sequences in which all 3 types of models are applicable.

Nucleotide substitution models define the relative occurrence of substitutions among the 4 nucleotides; hence, we refer to them as “4-dimensional” substitution models. Here, the “dimension” represents the number of states and the number of rows and columns of the transition rate matrix of the models. The simple Jukes–Cantor model (Jukes and Cantor 1969) assumes the equal occurrence of transitions and transversions and equal nucleotide frequencies, whereas the improved models relax either the former (Kimura 1980) or the latter assumptions (Felsenstein 1981). In addition, both assumptions can be relaxed simultaneously (Hasegawa et al. 1985), and the type of transitions can be distinguished (Tamura and Nei 1993). All these models are generalized by the general time-reversible model (GTR; Tavaré 1986; Yang 1994) in which the 4 nucleotide frequencies and 12 possible types of substitutions can be different. Amino acid substitution models such as JTT (Jones et al. 1992), Dayhoff (Dayhoff et al. 1978), mtREV (Adachi and Hasegawa 1996), WAG (Whelan and Goldman 2001), and so on, define the relative occurrence of substitutions among the 20 amino acids; thus, we refer to them as “20-dimensional” substitution models. Database-derived information for substitution patterns is obtained by using parsimony-based methods (e.g., Dayhoff et al. 1978; Jones et al. 1992) or likelihood-based methods (e.g., Adachi and Hasegawa 1996; Whelan and Goldman 2001). In the simple “proportional amino acid model” (Cao et al. 1994), all types of amino acid substitutions are equally probable, and the mechanistic amino acid models that consider the mutational biases of nucleotides (Yang et al. 1998) are also available. The amino acid frequencies can be derived either from a database or from the analysis of sequence data. In case the latter technique is used, the models have been found to be considerably improved; the names of such models are assigned the suffix “+F” (Cao et al. 1994). Codon substitution models define the relative occurrence of substitutions among sense codons (Goldman and Yang, 1994; Muse and Gaut, 1994). For simplicity, we describe codon models using the universal codon table; hence, we refer to them as “61-dimensional” substitution models. However, the specific number of dimensions of a codon model is not crucial in our study, and our descriptions are equally valid in codon models that use different codon tables. Using the codon model developed by Goldman and Yang (1994; GY94), the instantaneous transition rate from codon $i$ to codon $j$ is defined as follows:

$$
Q^{(61)}_{ij} = \begin{cases} 
0, & \text{more than one DNA substitution} \\
\pi_i, & \text{synonymous transversion} \\
k\pi_j, & \text{synonymous transition} \\
\omega\pi_j, & \text{nonsynonymous transversion} \\
(\omega k\pi)_j, & \text{nonsynonymous transition}
\end{cases}
$$

(1)
where $\pi_j$ and $k$ represent the frequency of codon $j$ and the relative occurrence of transitions with respect to transversions, respectively, and $\omega$ represents the relative occurrence of nonsynonymous substitutions with respect to synonymous substitutions and is used to detect deviation from neutral evolution. The superscript “(61)” indicates that the transition is among 61 sense codons. The codon model in equation 1 can be improved by allowing instantaneous multiple nucleotide substitutions (Whelan and Goldman 2004) or by incorporating the physical properties of proteins (Robinson et al. 2003) and the empirical substitution pattern derived from a database (Doron-Faigenboim and Pupko 2007; Kosiol et al. 2007; Seo and Kishino 2008).

Because statistical models are “approximations” of natural phenomena, all statistical models are fundamentally inaccurate. Thus, it is very crucial to select an appropriate model that minimizes the discrepancy between the model and the true evolutionary process. It has been noted that the selection of inappropriate models may result in the incorrect estimation of phylogenetic relationships and the incorrect estimation of uncertainty of tree topologies (e.g., Gaut and Lewis 1995; Sullivan and Swofford 1997; Bruno and Halpern 1999; Buckley 2002; Anderson and Swofford 2004). In the Bayesian framework, incorrect branch-length estimation due to the adoption of inappropriate models may cause incorrect estimation of divergence time (Thorne et al. 1998) or incorrect estimation of the trend in selective pressure (Seo et al. 2004).

Typically, likelihood-based models are compared using both frequentist and Bayesian inference methods. When one model is nested within another, that is, when one model is a special case of the other model, the likelihood ratio test (LRT; Stuart and Ord 1996) can be adopted to test the significance of the advantages of nesting models. For various nucleotide substitution models, hierarchical LRT can be adopted to select the set of appropriate models (e.g., Frati et al. 1997; Sullivan et al. 1997; Posada and Crandall 2001). For models that are not nesting or nested, LRT is not applicable, and the Akaike information criterion (AIC; Akaike 1974) or the Bayesian information criterion (BIC; Schwarz 1974) can be adopted for model selection. For a given data set ($X$) comprising $n$ independent data items ($X = \{x_1, \ldots, x_n\}$) and model $i$, the AIC and BIC are defined as:

$$\text{AIC}(i) = -2 \log(\hat{\theta}_i | X) + 2K \quad (2)$$

$$\text{BIC}(i) = -2 \log(\hat{\theta}_i | X) + K \log(n) \quad (3)$$

where $\hat{\theta}_i$ is the vector of the maximum likelihood estimates of model $i$, $K$ is the dimension of $\hat{\theta}_i$, and $\log(\hat{\theta}_i | X)$ is the log-likelihood score for $\hat{\theta}_i$ and $X$. The model whose AIC (or BIC) score is the minimum score is regarded as the best model. Models can also be compared using a decision-theoretic approach (Minin et al. 2003; Abdo et al. 2005) and the Bayes factor (e.g., Suchard et al. 2001; Aris-Brosou and Yang 2002). In these comparisons, appropriate loss functions and posterior distributions have to be determined by using the likelihood calculated using substitution models.

Although considerable attention has been paid to the comparisons among the same types of models (e.g., Nielsen and Yang 1998; Posada and Crandall 1998; Yang and Nielsen 1998, 2002; Yang et al. 2000; Abascal et al. 2005), relatively little attention has been paid to the comparisons among different types of models (Shapiro et al. 2006; Seo and Kishino 2008). The comparisons among different types of models are difficult because these models have different data structures, and hence, the AIC or BIC cannot be applied in a straightforward manner in these comparisons. For calculating the AIC (or BIC) scores for different models, the data $X$ in equation 2 (or equation 3) should be identical. Although the entire data set of aligned sequences is identical, individual data items are different for different types of models. When applying nucleotide models, we separate each codon column into 3 nucleotide columns. This “codon separation” yields different data $X$ for nucleotide and codon models, and the number of individual data items for nucleotide models becomes $3n$. For applying amino acid models, we translate codons into amino acids. Even a “translation” yields different data $X$ for amino acid and codon models, although the number of individual data items is identical in this case.

Recently, we have developed a new procedure to statistically compare amino acid models with codon models (Seo and Kishino 2008). By converting the 20-dimensional amino acid models into 61-dimensional codon models and applying amino acid models, it is possible to measure the log-likelihood scores corresponding to translations. Further, amino acid models can be regarded as special cases of codon models, and both types of models with the same 61-dimensional data structure can be compared (see Theory). Although amino acid models can be compared with codon models, the technique to compare these models with 4-dimensional nucleotide models is still unclear. A comparison of the AIC scores calculated using the nucleotide and codon models has been attempted (Shapiro et al. 2006); however, no theoretical justification has been provided for the comparison. Here, we solve the problem of comparing nucleotide and codon models despite their different data structures. We propose new procedures to convert both 4-dimensional nucleotide models and 61-dimensional codon models into 64-dimensional codon models. In these 64-dimensional codon models, stop codons as well as sense codons are taken into consideration. Thus, 20-dimensional amino acid models can be converted initially into 61-dimensional codon models via our previous procedure (Seo and Kishino 2008) and then into 64-dimensional codon models using our new procedure. The resulting 64-dimensional codon models are mathematically equivalent to 4-dimensional nucleotide, 20-dimensional amino acid, and 61-dimensional codon models. Furthermore, due to this mathematical equivalence, it is possible to obtain identical data $X$ in equations 2 and 3 and to compare the 3 types of models using the same 64-dimensional data structure.
structure. We have found that it is always possible to define a codon model, which is better than the given nucleotide model, by converting 4-dimensional nucleotide models into 64-dimensional codon models. It is always possible to define a codon model, which is better than the given amino acid model, by converting 20-dimensional amino acid models into 61-dimensional codon models (Seo and Kishino 2008). These theoretical findings exhibit the potential advantages and applications of codon models. We have applied our procedures to the analysis of 1) highly divergent interspecific mammalian mitochondrial, 2) highly conserved interspecific mammalian nuclear, and 3) polymorphic intraspecific human mitochondrial protein–coding sequences. Using empirical data analysis, we demonstrate and discuss the advantages of codon models over the other 2 types of models.

**Theory**

To begin with, we provide an overview of our previous procedure (Seo and Kishino 2008) used to convert a 20-dimensional amino acid model into 61-dimensional codon models and then explain the new procedures used to convert nucleotide and codon models into 64-dimensional models. We demonstrate our conversion procedures using the HKY (Hasegawa et al. 1985) and GY94 (Goldman and Yang 1994) models, but these conversion procedures can be easily applied to any type of nucleotide or codon model.

**Converting 20-Dimensional Amino Acid Models into 61-Dimensional Codon Models: Overview**

Let us consider an amino acid model in which the instantaneous transition rate is defined as

\[
R_{ij} = \begin{cases} 
    s_{a_i a_j} \pi_i \pi_j, & \text{if } a_i \neq a_j \\
    - \sum_{a_i' a_j' \neq a_i} R_{a_i' a_j'} \pi_i \pi_j, & \text{if } a_i = a_j 
\end{cases} 
\]  

(4)

where \( \pi_j \) is the frequency of the amino acid \( a_j \) and \( s_{a_i a_j} \) represents the relative occurrence of substitutions from \( a_i \) to \( a_j \) during an infinitesimal time period. The \( s_{a_i a_j} \) values are typically estimated by parsimony (Dayhoff et al. 1978; Jones et al. 1992) or likelihood (Adachi and Hasegawa 1996; Whelan and Goldman 2001) methods. From the amino acid model in equation 4, we derive a 61-dimensional SK-P1 model (Seo and Kishino 2008) in which the instantaneous transition rate from codon \( i \) to codon \( j \) is defined as

\[
R_{ij} = \begin{cases} 
    s_{a_i a_j} \pi_j, & \text{nonsynonymous change} \\
    \rho \pi_i \pi_j, & \text{synonymous change}
\end{cases} 
\]  

(5)

where \( a_i \) is the amino acid that codon \( j \) encodes and \( \pi_j \) is the frequency of codon \( j \). The \( s_{a_i a_j} \) values are obtained from the amino acid model in equation 4, and the free parameter \( \rho \) represents the relative occurrence of synonymous substitutions with respect to \( s_{a_i a_j} \). The transition probabilities during time \( t \) in the SK-P1 model in equation 5 can be analytically obtained using the transition probabilities and eigenvalues of the amino acid model in equation 4. When \( \rho = \infty \) in equation 5, synonymous substitutions are completely saturated, and the SK-P1 model is reduced to an “SK-P0” model, which is mathematically equivalent to the amino acid model in equation 4.

The empirical \( s_{a_i a_j} \) values obtained from equation 4 can be used to improve the conventional codon model in equation 1; the improved model is referred to as the “SK-P2” model (Seo and Kishino 2008). For the SK-P2 model, the instantaneous transition rate from codon \( i \) to codon \( j \) is defined as

\[
R_{ij} = \begin{cases} 
    0, & \text{more than one DNA substitution} \\
    \rho \pi_i \pi_j, & \text{synonymous transversion} \\
    \rho \kappa \pi_i \pi_j, & \text{synonymous transition} \\
    s_{a_i a_j} / \pi_j, & \text{nonsynonymous transversion} \\
    s_{a_i a_j} \kappa \pi_i \pi_j, & \text{nonsynonymous transition}
\end{cases} 
\]  

(6)

When \( s_{a_i a_j} = 1 \) and \( \rho = 1 / \omega \), the SK-P2 model becomes identical to the GY94 model. That is, the GY94 model is an SK-P2 model in which the simple proportional amino acid model \( (s_{a_i a_j} = 1; \text{Cao et al. 1994}) \) is incorporated. Because any amino acid model can be incorporated into the SK-P2 model, the SK-P2 model is more flexible than the GY94 model.

**Converting 4-Dimensional HKY Model into 64-Dimensional Model**

In the HKY model (Hasegawa et al. 1985), the instantaneous transition rate from nucleotide \( a \) to nucleotide \( b \) is defined as follows:

\[
R_{ab}^{(4)} = \begin{cases} 
    \kappa \pi_b, & \text{transition} \\
    \pi_a, & \text{transversion}
\end{cases} 
\]  

(7)

where \( \pi_b \) represents the frequency of nucleotide \( b \). We use the superscript “(4)” to indicate that equation 7 defines the instantaneous transition rates among the 4 nucleotides. Let us denote the \( 4 \times 4 \) rate matrix derived from equation 7 as \( R^{(4)} \). Then, the transition probability from nucleotide \( a \) to nucleotide \( b \) during time \( t \) is given by the element in the \( a \)th row and \( b \)th column of \( P^{(4)}(t) = \exp(t R^{(4)}) \). For simplicity, we do not consider the normalizing factors of rate matrices in our study.

Similar to the definition in equation 7, we define a 64-dimensional codon model in which the instantaneous transition rate from codon \( r \) to codon \( s \) is given as follows:

\[
R_{rs}^{(64)} = \begin{cases} 
    0, & \text{more than one DNA substitution} \\
    \kappa \pi_s, & \text{transition occurs at } p \text{th site} \\
    \pi_r, & \text{transversion occurs at } p \text{th site}
\end{cases} 
\]  

(8)

where \( p \in \{1, 2, 3\} \) and \( \pi_p \) is the nucleotide of codon \( s \) at the \( p \)th site (refer Whelan and Goldman [2004] for a
similar conversion procedure within 61 dimensions). Let us denote the 64 × 64 rate matrix derived from equation 8 as \( R^{(64)} \). Then, the transition probability from codon \( r \) to codon \( s \) during time \( t \) is given by the element in the \( r \)th row and \( s \)th column of \( P^{(64)}(t) = \exp(tR^{(64)}) \).

We have found that the transition probabilities of the \( R^{(64)} \) model can be represented by multiplication of the transition probabilities of the \( R^{(4)} \) model at each nucleotide site (see Appendix). That is,

\[
p_{rs}(t) = p_{r1s1}(t) \cdot p_{r2s2}(t) \cdot p_{r3s3}(t).
\]

Further, we have found that the likelihood of the phylogeny obtained using the \( R^{(64)} \) model is identical to that obtained using the \( R^{(4)} \) model. Let us denote the protein-coding sequence data as \( S = s_1s_2 \cdots s_n \), where \( s_i \) is the \( i \)th column of nucleotide alignment and \( n \) is the number of codon columns. Let \( \theta \) denote the unknown parameters of the rate matrix and the time duration along the branches of phylogeny. Then, for the \( i \)th column of codon alignment, we can obtain the following relationship (see Appendix):

\[
L^{(64)}(\theta|s_{3i-2}s_{3i-1}s_{3i}) = L^{(4)}(\theta|s_{3i-2}) \cdot L^{(4)}(\theta|s_{3i-1}) \cdot L^{(4)}(\theta|s_{3i}),
\]

where \( L^{(64)}(\theta|\cdot) \) and \( L^{(4)}(\theta|\cdot) \) represent the likelihoods calculated using the \( R^{(64)} \) and \( R^{(4)} \) models, respectively. Using equation 10, we obtain the following relationship:

\[
L^{(64)}(\theta|S) = \prod_{i=1}^{n} L^{(64)}(\theta|s_{3i-2}s_{3i-1}s_{3i})
\]

\[
= \prod_{i=1}^{n} \left[ L^{(4)}(\theta|s_{3i-2}) \cdot L^{(4)}(\theta|s_{3i-1}) \cdot L^{(4)}(\theta|s_{3i}) \right]
\]

\[
= \prod_{i=1}^{n} L^{(4)}(\theta|s_i) = L^{(4)}(\theta|S).
\]

That is, the likelihood calculated using the nucleotide model \( R^{(4)} \) is identical to that calculated using the codon model \( R^{(64)} \). This implies that we can regard a 4-dimensional nucleotide model as a 64-dimensional codon model in which the nucleotides at the first-, second-, and third-codon sites are grouped together as a single unit in the evolutionary process.

The instantaneous transition rate given by equation 8 can be nonzero even when \( r \) and/or \( s \) are stop codons. This allows nucleotide substitutions to occur unrestrictedly and independently at the 3 codon sites; this independence is consistent with the factorization of equations 9 and 10. Because of the independence among the 3 codon sites in the \( R^{(64)} \) model, the stationary frequency \( \pi_r \) is given by \( \pi_r = \pi_1 \pi_2 \pi_3 \), even when \( r \) is a stop codon. The nonzero instantaneous transition rates when a stop codon is (or stop codons are) involved and the nonzero frequencies of stop codons are the 2 main causes of the poor performance of nucleotide models (see Results and Discussion).

In general, the third-codon sites in protein-coding sequences evolve faster than the first- or second-codon sites because most nucleotide substitutions at the third-codon sites are synonymous. Thus, typically, an additional free parameter is assigned to the third-codon sites to allow a high evolutionary rate in the analysis using 4-dimensional nucleotide models. This type of nucleotide model can also be converted into a 64-dimensional codon model. More generally, we can assign different evolutionary rates for the 3 codon sites. In equation 8, if we multiply \( \alpha \), \( \beta \), and \( \gamma \) with the instantaneous rates during nucleotide substitutions at the first-, second-, and third-codon sites, respectively, then equation 9 changes into

\[
p_{rs}(t) = p_{r1s1}(\alpha t) \cdot p_{r2s2}(\beta t) \cdot p_{r3s3}(\gamma t),
\]

which can be easily proven by modifying the proof in the Appendix. By setting \( \alpha = \beta = 1 \) and \( \gamma \) as a free parameter, we can define an \( R^{(64)} \) model equivalent to an \( R^{(4)} \) model in which the third-codon sites are fast-evolving sites. Furthermore, by assuming that \( \alpha, \beta, \) and \( \gamma \) follow the gamma distribution, we can convert the nucleotide model with gamma rate heterogeneity (Yang 1993) into a 64-dimensional model (see Results and Discussion).

### Converting a 61-Dimensional Codon Model into a 64-Dimensional Codon Model

This conversion procedure is rather straightforward. For a given 61-dimensional codon model, we consider a 64-dimensional codon model in which the instantaneous transition rate from codon \( r \) to codon \( s \) is

\[
Q^{(64)}_{rs} = \begin{cases} Q^{(61)}_{rs} & \text{if } r \text{ and } s \text{ are sense codons} \\ 0 & \text{otherwise} \end{cases},
\]

where \( Q^{(61)}_{rs} \) is the instantaneous transition rate from codon \( r \) to codon \( s \) in the given 61-dimensional codon model, as in equations 1, 5, and 6. Then, the 64 × 64 rate matrix derived from equation 12 is represented as

\[
Q^{(64)} = \begin{pmatrix} Q^{(61)} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},
\]

\[
\exp(tQ^{(64)}) = \begin{pmatrix} \exp(tQ^{(61)}) & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.
\]
using the $Q^{(64)}$ model are identical to those obtained using the $Q^{(61)}$ model. Second, stop codons are explicitly considered, but transitions from a sense codon (or from a stop codon) to a stop codon (or to a sense codon) never occur, and stop codons, if observed in protein-coding sequences, remain constant. In general, stop codons are not observed in protein-coding sequences, and we can adopt the sense codon frequencies of the $Q^{(61)}$ model for the $Q^{(64)}$ model. Then, the likelihoods of the $Q^{(61)}$ and $Q^{(64)}$ models are identical, and the $Q^{(61)}$ model can be regarded as the $Q^{(64)}$ model. In terms of the likelihood scores, there is no change in the $Q^{(64)}$ model by explicitly considering stop codons.

**Statistical Comparison of 3 Types of Models**

Using our conversion procedures, we have theoretically shown that 4-dimensional nucleotide models and 61-dimensional codon models can be regarded as 64-dimensional codon models, and their likelihoods can be directly compared. In our previous study (Seo and Kishino 2008), we showed that the likelihoods of 20-dimensional amino acid models and 61-dimensional codon models are not directly comparable. Because the likelihood corresponding to translations is explicitly considered in the SK-P0 model, amino acid models are comparable to codon models only through SK-P0 models. Therefore, the likelihoods of nucleotide, SK-P0, and codon models are comparable, and the corresponding AIC or BIC scores can be adopted to determine the best model among the 3 types.

Our procedure for the conversion of a 4-dimensional nucleotide model into a 64-dimensional codon model clearly implies that the BIC scores should be calculated carefully within the type of nucleotide models. For the sequence data of $n$ aligned codon columns, the number of samples in the simple nucleotide models is $3n$; in these models, the 3 codon sites are assumed to be independently and identically distributed. Therefore, in equation 3, the following equation should be adopted for calculating the BIC scores:

$$\text{BIC}(i) = -2l(\hat{\theta}_i|X) + K \log(3n).$$

However, when the 3 codon sites are assumed to evolve differently, the 3 consecutive nucleotide sites should be grouped together and regarded as a single unit of sequence evolution. Therefore, equation 3 should be adopted for calculating the BIC scores even for the comparison of nucleotide models.

**RESULTS AND DISCUSSION**

**Sequence Data**

For the comparisons of the 3 types of models, we analyzed 3 sequence data sets comprising 1) fast-evolving interspecific, 2) slowly evolving interspecific, and 3) highly polymorphic intraspecific protein-coding sequences.

The first data set consisted of 12 mitochondrial genes obtained from 69 mammalian species (Nikaido et al. 2003). The data sequences were downloaded from the GenBank sequence database (for accession number, see Nikaido et al. 2003), aligned using ClustalW (Chenna et al. 2003) with the default settings, and manually edited. The log-likelihood scores for various models were calculated using “Tree 6” estimated by Nikaido et al. (2003).

For the second data set, we selected the 10 most slowly evolving genes from the 2789 nuclear genes of 10 mammals that were analyzed by Nishihara et al. (2007) to investigate the phylogenetic relationship among the Boreoeutheria, Xenartha, and Afrotheria groups. Using the phylogenies estimated by Nishihara et al. (2007), we calculated the likelihood scores of 10 genes using various models.

The third data set consisted of 12 mitochondrial genes of humans. We excluded individuals from the same location and selected 33 human mitochondrial genomes from the data set obtained by Ingman et al. (2000) for the investigation of human origin. The mitochondrial genomes were separated into 12 protein-coding sequences, which were aligned using ClustalW (Chenna et al. 2003) with the default settings and then manually edited. The log-likelihood scores for various models were calculated using the neighbor-joining (Saitou and Nei 1987) tree topology estimated by Ingman et al. (2000).

In the analysis of the 3 data sets, we have not considered the uncertainty in the tree topologies and have assumed that the adopted topologies represent the true phylogenetic relationships. However, we expect that our results are robust for different tree topologies because model selection has been known to be rarely affected by tree topologies if the adopted topologies are not considerably different from the true topologies (Sullivan and Swofford 1997; Posada and Crandall 2001; Abdo et al. 2005).

**Three Types of Models and Their Conversion to 64-Dimensional Models**

**DNA model and conversion.**—Although our conversion procedure is generally applicable for any type of nucleotide model, we demonstrate it here using the HKY model (Hasegawa et al. 1985). To explicitly denote that equation 8 is derived from the HKY model in equation 7, we will refer to the 64-dimensional model in equation 8 as the HKY($^{(64)}$) model in our analysis. We can make the HKY($^{(64)}$) model more realistic by assigning an additional free parameter in the rate matrix given by equation 8 during the substitutions at the third-codon sites and we refer to this new model as the HKY($^{(64)}$)/3rd model. Further, the HKY($^{(64)}$)/3rd model can be made more realistic by assigning zero instantaneous transition rates when stop codons are involved and refer to this new model as the HKY($^{(61)}$)/3rd model. While there exist 4-dimensional nucleotide models equivalent to the HKY($^{(61)}$)/3rd model and HKY($^{(64)}$)/3rd models, there exist no nucleotide models equivalent to the HKY($^{(61)}$)/3rd model.
model. Therefore, the HKY\(^{(61)}\)/3rd model should be categorized as a 61-dimensional codon model. The HKY\(^{(61)}\)/3rd model is a 61-dimensional codon model that is closest to the HKY model.

**Amino acid models and their conversion.**—Different amino acid models were adopted for different data sets. For the mammalian and human mitochondrial genes, we adopted the mREV + F (Adachi and Hasegawa 1996) amino acid model. For the mammalian nuclear genes, we adopted the WAG + F (Whelan and Goldman 2001) amino acid model. These amino acid models were converted into SK-P0 and SK-P1 models, respectively.

**Codon models and their conversion.**—As we noted, the 61-(or 60) dimensional codon models can be regarded as 64-dimensional codon models. For a statistical comparison of the 3 types of models, we considered 3 low-dimensional codon models: 1) GY94 (Goldman and Yang 1994), 2) SK-P1 (equation 5), and 3) SK-P2 (equation 6) models. The \(s_{ij}\) values of the mtREV (Adachi and Hasegawa, 1996) and WAG (Whelan and Goldman 2001) models were incorporated into the SK-P1 and SK-P2 models for the analysis of the mitochondrial and nuclear protein-coding sequences, respectively.

**Comparing Log-Likelihoods of 7 Different Models**

Because our selection of 7 different models and 3 data sets is rather arbitrary, some results of model comparison might be highly specific for the compared models and the analyzed data. However, these results are very useful in elucidating the advantages, disadvantages, and performance of the 3 types of models. Other results of the model comparison are so fundamental and general that they will be robust for selections of candidate models and sequence data. These results are discussed in detail in this section.

Tables 1–3 show the log-likelihood scores of the 7 different models for the 1) interspecific mammalian mitochondrial, 2) interspecific mammalian nuclear, and 3) intraspecific human mitochondrial protein-coding sequences, respectively. The number of free parameters in the rate matrices varies between 0 and 2. This range is considerably smaller than the range of differences between the log-likelihood scores of the 7 different models. For example, the longest gene in Table 1 is NADH5, whose length is 616 codons including gaps. The difference in \(K \log n\) between the models with \(K = 2\) and \(K = 0\) is 12.8; this difference is relatively smaller than the difference in the log-likelihood scores and has negligible effect on the AIC and BIC scores. Therefore, we list the log-likelihood scores in Tables 1–3 instead of the AIC or BIC scores.

The consistency in the results in Tables 1–3 indicates that codon models generally show better performances than amino acid (SK-P0) and nucleotide models (HKY\(^{(64)}\) and HKY\(^{(64)}\)/3rd). The SK-P2 and GY94 models are always better than the other models, and the HKY\(^{(61)}\)/3rd model is always better than the HKY\(^{(64)}\) and HKY\(^{(64)}\)/3rd

### TABLE 1. Maximum log-likelihood scores of the 7 different models for 12 mitochondrial genes from 69 mammalian species

<table>
<thead>
<tr>
<th>Gene</th>
<th>Model</th>
<th>AIC</th>
<th>BIC</th>
<th>AICn</th>
<th>BICn</th>
</tr>
</thead>
<tbody>
<tr>
<td>COX1</td>
<td>HKY(^{(61)})</td>
<td>75934</td>
<td>76489</td>
<td>75934</td>
<td>76489</td>
</tr>
<tr>
<td></td>
<td>HKY(^{(61)})/3rd</td>
<td>72989</td>
<td>73544</td>
<td>72989</td>
<td>73544</td>
</tr>
<tr>
<td></td>
<td>SK-P0 (equation 5)</td>
<td>73544</td>
<td>74100</td>
<td>73544</td>
<td>74100</td>
</tr>
<tr>
<td></td>
<td>SK-P1 (equation 5)</td>
<td>73544</td>
<td>74100</td>
<td>73544</td>
<td>74100</td>
</tr>
<tr>
<td></td>
<td>SK-P2 (equation 6)</td>
<td>73544</td>
<td>74100</td>
<td>73544</td>
<td>74100</td>
</tr>
<tr>
<td></td>
<td>GY94 (equation 6)</td>
<td>73544</td>
<td>74100</td>
<td>73544</td>
<td>74100</td>
</tr>
</tbody>
</table>

Notes: The number of parameters in the rate matrix is represented by \(p\).
models.

In all cases, in Tables 1–3, the log-likelihood scores of the HKY(61)/3rd model are greater than those of the HKY(64) model. This observation is consistent with the fact that the third-codon sites are mostly synonymous sites, and synonymous substitutions generally occur more often than nonsynonymous substitutions. The advantages of the HKY(61)/3rd model over the HKY(64)/3rd model can be explained by the effect of stop codons. As we noted in Theory, nonzero frequencies of the stop codons and nonzero instantaneous transition rates when a stop codon is (or stop codons are) involved are explicitly assumed in the HKY(64)/3rd model. These unrealistic assumptions for protein-coding sequences are intentionally excluded in the HKY(61)/3rd model, which leads to the improvement of the log-likelihood scores. This provides strong evidence of the superiority of codon models, and the superiority of codon models can be generalized for different nucleotide models. For any given 4-dimensional nucleotide model, we can always define a 61-dimensional codon model in which the above-mentioned unrealistic assumptions are excluded. That is, there exists at least one codon model that is better than the given nucleotide model. Similarly, consider the GTR(61)/3rd, GTR(64)/3rd, and GTR(64) models, which are derived from 4-dimensional GTR models (Tavaré 1986; Yang 1994). Like the HKY(61)/3rd model, the GTR(61)/3rd model explicitly excludes stop codons and is expected to be better than the GTR(64)/3rd and GTR(64) models. Because there are 4 additional free parameters in the rate matrix, the GTR(61)/3rd model requires more computation than the HKY(61)/3rd model and may not be practical for the analysis of a large number of genes and taxa. However, our fundamental and general conclusion will still hold in this case; it is always possible to define a codon model that might be computationally impractical but will be better than the given nucleotide model.

In Table 1, the log-likelihood scores of the SK-P0 model are generally greater than those of the HKY(61)/3rd model, except in the case of ATP8, whose length is 70 codons including gaps. Due to its short sequence length, ATP8 may not provide sufficient resolution for model comparison. Divergent protein-coding sequences contain a reasonable number of amino acid substitutions, and the substitution pattern will be consistent with the $s_{a,a_i}$ values in equation 4. In Table 1, the additional information provided by the $s_{a,a_i}$ values indicates the superiority of the SK-P0 model over the HKY(61)/3rd model. However, when the sequences are not divergent (Tables 2 and 3), the observed pattern of amino acid substitutions will be different from the empirical $s_{a,a_i}$ values. In this case, the HKY(61)/3rd model in which only mechanistical nucleotide substitutions are considered is better than the SK-P0 model.

The advantages of the SK-P1 model over the SK-P0 model have been theoretically investigated in our
TABLE 3. Maximum log-likelihood scores of the 7 different models for 12 mitochondrial genes from 33 humans

<table>
<thead>
<tr>
<th>Gene</th>
<th>Model</th>
<th>sHK</th>
<th>sHK/3rd</th>
<th>sHK/3rd</th>
<th>sHK/3rd</th>
<th>sHK/3rd</th>
<th>sHK/3rd</th>
<th>sHK/3rd</th>
<th>sHK/3rd</th>
</tr>
</thead>
<tbody>
<tr>
<td>COX1</td>
<td>HKY(61)</td>
<td>150</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
</tr>
<tr>
<td>COX2</td>
<td>HKY(61)</td>
<td>150</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
</tr>
<tr>
<td>COX3</td>
<td>HKY(61)</td>
<td>150</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
</tr>
<tr>
<td>NADH4</td>
<td>HKY(61)</td>
<td>150</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
</tr>
<tr>
<td>NADH5</td>
<td>HKY(61)</td>
<td>150</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
</tr>
<tr>
<td>NADH6</td>
<td>HKY(61)</td>
<td>150</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
</tr>
<tr>
<td>ATP6</td>
<td>HKY(61)</td>
<td>150</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
</tr>
<tr>
<td>ATP8</td>
<td>HKY(61)</td>
<td>150</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
<td>150/3rd</td>
</tr>
</tbody>
</table>

Adapted from Seo and Kishino (2008). We have observed the same results for protein-coding sequences that are not divergent (Tables 2 and 3). The superiority of the SK-P1 model over the SK-P0 model can be generalized, and it is always possible to define a 61-dimensional SK-P1 model that is better than the given amino acid model (Seo and Kishino 2008). In Tables 2 and 3, the SK-P0 model shows very poor performance, and its log-likelihood scores are remarkably lower than those of the HKY(61) model. When sequences are not divergent, the SK-P0 model (equivalent amino acid model) is not applicable because of the small number of amino acid substitutions.

Except for ATP8 in Table 1 and EPC2 in Table 2, the log-likelihood scores of the SK-P1 model are greater than those of the HKY(61)/3rd model, implying that the SK-P1 model is generally better than the HKY(61)/3rd model. Because the length of EPC2 is 191 codons, this exception may not be caused by the small length. The value of k for EPC2 estimated using the HKY(61)/3rd model is 4.11. Thus, for EPC2, it appears that separating transitions and transversions is more realistic than incorporating the empirical information of the WAG + F (Whelan and Goldman 2001) amino acid model.

Multiple nucleotide substitutions are allowed in the SK-P1 model during an infinitesimal time period, whereas only a single nucleotide substitution is allowed in the SK-P2 model. Further, transitions and transversions are distinguished in the SK-P2 model. A single transition and a single transversion during an infinitesimal time period appear to be consistent with the real evolutionary mechanism; this consistency explains the reason behind the log-likelihood scores of the SK-P2 model being greater than those of the SK-P1 models in all cases in Tables 1–3. In Table 1, the SK-P2 model is better than the GY94 model, except for the case of ATP8, which is due to the empirical information provided by the s_{adj} values. Although the GY94 and SK-P2 models show overall superiority in Tables 2 and 3, respectively, the differences in log-likelihood scores are relatively small. Hence, we expect that these two models may be equally applicable for conserved protein-coding sequences.

Advantages and Disadvantages of 3 Types of Models and Concluding Remarks

Using our conversion procedures, it is possible to statistically compare 4-dimensional nucleotide models with 61-dimensional codon models. We have shown that there exists at least one 61-dimensional codon model that is better than the given nucleotide model. Using our previous procedure to convert a 20-dimensional amino acid model into a 61-dimensional codon model (Seo and Kishino 2008), we can find at least one 61-dimensional codon model that is better than the given amino acid model. Therefore, we note that a codon model is always the best model among evolutionary models. That is, we can always define at least one codon model that is better
than the given nucleotide and amino acid models.

Although codon models are in general better than the other 2 types of models, they have a disadvantage—they require intensive computation. When optimizing the free parameters of a rate matrix, we must recalculate the eigenvalues and eigenvectors of a 61-dimensional rate matrix for updated free parameters. Furthermore, codon models require intensive computation for the pruning algorithm (Felsenstein 1981) because of their high dimensions. In contrast, 4-dimensional nucleotide models require considerably less computation. The eigenvalues, eigenvectors, and transition probabilities are analytically available in many cases, and the pruning algorithm is quite fast. The computational advantage of 4-dimensional nucleotide models over the other 2 types of modes is more apparent when we consider the rate heterogeneity among sites (RHAS; Yang 1993). However, nucleotide models are expected to show poor performance, as represented in Tables 1–3. For divergent protein-coding sequences, we note that highly sophisticated nucleotide models that consider RHAS can be worse than the simple SK-P1 model that does not consider RHAS. For the 12 protein-coding sequences listed in Table 1, we have calculated the maximum log-likelihood scores by using PAML (Yang 2007) and the GTR(64)/3rd model with gamma distribution of RHAS, which is referred to as the GTR(64)/3rd+Γ model. Except NADH5 and ATP8, the other 10 genes show better performance by the application of the SK-P1 or SK-P2 model as compared with the application of the GTR(64)/3rd+Γ model (data not shown). The behavior of ATP8 can be explained by the lack of resolution caused by its small sequence length, and the behavior of NADH5 can be explained by 2 factors: consideration of RHAS and inclusion of additional free parameters in the rate matrix of the GTR(64)/3rd+Γ model. Because these 2 factors require intensive computation, only one of them may be practically incorporated into codon models. As an example, we have calculated the maximum log-likelihood score of NADH5 using the GY94 model and RHAS (GY94 + Γ) using PAML (Yang 2007). We have observed that the GY94 + Γ model performs much better than the GTR(64)/3rd+Γ model in the analysis of NADH5 (data not shown), which is considered to be due to the explicit exclusion of stop codons by using the GY94 + Γ model.

Amino acid models are intermediate between nucleotide models and codon models in terms of their computational load. The computational load for pruning algorithms is heavier than that of nucleotide models, but lighter than that of codon models. Further, amino acid models do not have free parameters in the rate matrix, and the recalculation of the eigenvalues and eigenvectors is not required. However, amino acid models are expected to show poor performance when the sequences are not divergent (Tables 2 and 3).

When a large number of genes and taxa have to be analyzed and the computational load is an important issue, our new codon model, the SK-P1 model, can be a good alternative to nucleotide models and other codon models. Although the computational load for the pruning algorithm is the same as that of a conventional 61-dimensional codon model, the eigenvalues and eigenvectors need not be recalculated for optimizing the ρ parameter. Further, a 61-dimensional transition probability matrix can be analytically derived from the 20-dimensional transition probability matrix of an amino acid model (Seo and Kishino 2008), drastically reducing the computational time. Furthermore, the SK-P1 model retains the most important merit of codon models—synonymous and nonsynonymous substitutions can be separated and compared to detect the deviation from neutral evolution. When a small number of genes and taxa have to be analyzed or the computational load is not a major concern, our SK-P2 model can be a good alternative to the GY94 model because it has more flexibility for incorporating the empirical information of amino acid models.

In this paper, we have presented the theoretical foundation for the statistical comparison of nucleotide and codon models. In conjunction with our previous results (Seo and Kishino 2008), we conclude that codon models should be adopted for the evolutionary analysis of protein-coding sequences so long as the computational load is not a major concern. Further, with continuous improvements in computational performance in the postgenomic era, the development of realistic models of sequence evolution is becoming increasingly more important. Our conclusion on the superiority of codon models will provide a useful guide for future research on improving evolutionary models.

FUNDING

T.-K.S. was supported by JSPS (EYS B-18770216), and H.K. was supported by JSPS (SR B-16300086).

ACKNOWLEDGMENTS

We thank Jack Sullivan, Marc Suchard, David Posada, and an anonymous reviewer for providing valuable comments and Hidenori Nishihara for providing the sequence data.

REFERENCES


Received 7 June 2008; revised 26 August 2008; accepted 5 January 2009

APPENDIX

Proof of Equation 9. The matrix \( R^{(64)} \) of equation 8 can be separated into three matrices: \( R^{(64,1)} \), \( R^{(64,2)} \), and \( R^{(64,3)} \). For \( p = 1, 2, 3 \), the components of \( R^{(64, p)} \) are

\[
R_{rs}^{(64,p)} = \begin{cases} 
\frac{\kappa T r_{i}}{\tau_{o}}, & r = s \\
\pi_{s}, & r \neq s \ \\
\sum_{t \neq r} R_{st}^{(64, p)} - R_{rr}^{(64, p)}, & r = s \\
0, & \text{otherwise} 
\end{cases}
\]

transcription occurs only at pth site

transversion occurs only at pth site
Then,
$$R^{(64)} = R^{(64, 1)} + R^{(64, 2)} + R^{(64, 3)}.$$ 

We found that the three matrices commute with each other. That is, for $i$ and $j$ ($i \neq j$ and $i, j \in \{1, 2, 3\}$), $R^{(64, i)} R^{(64, j)} = R^{(64, j)} R^{(64, i)}$. This leads to (Golub and Van Loan, p. 577)

$$\exp\{t R^{(64)}\} = \exp\{t (R^{(64, 1)} + R^{(64, 2)} + R^{(64, 3)})\} = \exp\{t R^{(64, 1)}\} \exp\{t R^{(64, 2)}\} \exp\{t R^{(64, 3)}\}.$$ 

To calculate $\exp\{t R^{(64)}\}$, we separately calculate $\exp\{t R^{(64, p)}\}$ ($p = 1, 2, 3$). For $R^{(64, p)}$, we consider the operation of exchanging the $a$th and $b$th rows and columns, which is denoted by the matrix function $T_{a,b}[]$. $T_{a,b}[]$ simply changes the order of codons and does not affect the validity of the rate matrix. We found that iteratively exchanging the rows and columns transforms $R^{(64, p)}$ into a diagonal block matrix. That is, for each $p (= 1, 2, 3)$, we can find a set of integer pairs $\{(a_1^{(p)}, b_1^{(p)}), (a_2^{(p)}, b_2^{(p)}), \ldots, (a_4^{(p)}, b_4^{(p)})\}$ such that

$$T_{a_1^{(p)} b_1^{(p)}} \circ T_{a_2^{(p)} b_2^{(p)}} \circ \cdots \circ T_{a_4^{(p)} b_4^{(p)}} \{t R^{(64, p)}\} = \begin{pmatrix} \exp\{t R^{(4)}\} & 0 & \cdots & 0 \\ 0 & \exp\{t R^{(4)}\} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp\{t R^{(4)}\} \end{pmatrix}.$$ 

(15)

where $R^{(4)}$ is the rate matrix of equation 7 and 0 is a $4 \times 4$ null matrix. The operator $\circ$ denotes the composition of functions, that is, $T_{a_1^{(p)} b_1^{(p)}} \circ T_{a_2^{(p)} b_2^{(p)}} \circ \cdots \circ T_{a_4^{(p)} b_4^{(p)}} \{t R^{(64, p)}\} = T_{a_1^{(p)} b_1^{(p)}} T_{a_2^{(p)} b_2^{(p)}} \cdots T_{a_4^{(p)} b_4^{(p)}} \{t R^{(64, p)}\}.$

The exponentiation of equation 15 can be obtained by the exponentiation of diagonal blocks:

$$\exp\left\{ T_{a_1^{(p)} b_1^{(p)}} \circ T_{a_2^{(p)} b_2^{(p)}} \circ \cdots \circ T_{a_4^{(p)} b_4^{(p)}} \{t R^{(64, p)}\} \right\} = \begin{pmatrix} \exp\{t R^{(4)}\} & 0 & \cdots & 0 \\ 0 & \exp\{t R^{(4)}\} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \exp\{t R^{(4)}\} \end{pmatrix}.$$ 

(16)

Because $T_{a,b}[]$ simply changes the order of codons, applying $T_{a,b}[]$ to $R^{(64, p)}$ followed by exponentiation and applying $T_{a,b}[]$ again is identical to the exponentiation of $R^{(64, p)}$:

$$T_{a,b} \{\exp\{T_{a,b} \{t R^{(64, p)}\}\}\} = \exp\{t R^{(64, p)}\}.$$ 

In a similar way, applying $T_{a_1^{(p)} b_1^{(p)}} \circ \cdots \circ T_{a_4^{(p)} b_4^{(p)}}$ to equation 16 is identical to the exponentiation of $R^{(64, p)}$:

$$\begin{pmatrix} P^{(4)}(t) & 0 & \cdots & 0 \\ 0 & P^{(4)}(t) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & P^{(4)}(t) \end{pmatrix}.$$ 

where

$$P_{rs}^{(64, p)}(t) = \begin{cases} P_{r=sp}^{(4)}(t), & \text{substitution occurs only at } p\text{th site} \\ P_{r=sp}^{(4)}(t), & r = s \\ 0, & \text{otherwise} \end{cases}.$$ 

Let us denote $\exp\{t R^{(64, 1)}\}$ as $P^{(64, 1 \times 2)}(t)$. Then, the elements of $P^{(64, 1 \times 2)}(t)$ are

$$P_{r,s}^{(64, 1 \times 2)}(t) = \sum_{v=1}^{64} P_{r,v}^{(64, 1)}(t) P_{v,s}^{(64, 2)}(t).$$

The transition probability matrix derived from equation 8—$P^{(64)}(t)$—can be obtained with $\exp\{t R^{(64, 1 \times 2)}\} \times \exp\{t R^{(64, 3)}\}$, whose elements are

$$P_{r,s}^{(64)}(t) = \sum_{v=1}^{64} P_{r,v}^{(64, 1 \times 2)}(t) P_{v,s}^{(64, 3)}(t) = P_{r,s}^{(4)}(t) P_{r,s}^{(4)}(t) P_{r,s}^{(4)}(t).$$

Proof of equation 11. We prove equation 11 for a phylogeny of 4 taxa using 1 codon site. The proof can
straightforwardly be extended to the case with more than 4 taxa and the case of \( n \) codons. Suppose that we observe 4 terminal codons \( p = (p_1, p_2, p_3) \), \( q = (q_1, q_2, q_3) \), \( r = (r_1, r_2, r_3) \), and \( s = (s_1, s_2, s_3) \), where \( P_k, P_k', P_k'', \) and \( s_k (k = 1, 2, 3) \) are nucleotides. The unknown ancestral codon of \( p \) and \( q \) (\( r \) and \( s \)) is \( x = (x_1, x_2, x_3) \) \((y = (y_1, y_2, y_3))\). The time duration from \( x \) to \( p \) (\( q \)) is \( t_1 \) \((t_2)\) and the time duration from \( y \) to \( r \) (\( s \)) is \( t_3 \) \((t_4)\). The time duration between \( x \) and \( y \) is \( t_5 \). Then, the likelihood with a DNA model is

\[
L (4)(\theta | p, q, r, s)
\]

\[
= L (4)(\pm b\theta | p_1, q_1, r_1, s_1) \times L (4)(\theta | p_2, q_2, r_2, s_2)
\]

\[
\times L (4)(\theta | p_3, q_3, r_3, s_3)
\]

\[
= \sum_{x_1=1}^{4} \pi_{x_1} \sum_{x_2=1}^{4} \pi_{x_2} \sum_{y_1=1}^{4} \pi_{y_1} \sum_{y_2=1}^{4} \pi_{y_2}
\]

\[
\times (t_1) P(x_1, t_1) P(x_2, t_2) P(y_1, t_3) P(y_2, t_4)
\]

\[
\times \frac{4}{\sum_{x_3=1}^{4} \sum_{x_4=1}^{4} \sum_{y_3=1}^{4} \sum_{y_4=1}^{4} \pi_{x_3} \pi_{x_4} \pi_{y_3} \pi_{y_4} P(x_3, t_1) P(x_4, t_2) P(y_3, t_3) P(y_4, t_4)}
\]

\[
= \sum_{x_1=1}^{64} \pi_{x_1} \sum_{x_2=1}^{64} \pi_{x_2} \sum_{y_1=1}^{64} \pi_{y_1} \sum_{y_2=1}^{64} \pi_{y_2}
\]

\[
\times (t_1) P(x_1, t_1) P(x_2, t_2) P(y_1, t_3) P(y_2, t_4)
\]

\[
\times \frac{64}{\sum_{x_3=1}^{64} \sum_{x_4=1}^{64} \sum_{y_3=1}^{64} \sum_{y_4=1}^{64} \pi_{x_3} \pi_{x_4} \pi_{y_3} \pi_{y_4} P(x_3, t_1) P(x_4, t_2) P(y_3, t_3) P(y_4, t_4)}
\]

\[
= L (64)(\theta | p, q, r, s).
\]