A Codon Substitution Model with the Simultaneous Changes of Multiple Nucleotides for Protein Evolution

Sanzo Miyazawa

miyazawa@smlab.sci.gunma-u.ac.jp

Graduate School of Engineering, Gunma University, Japan

presented at

The 46th annual meeting of Biophysical Society of Japan in 2008 in Fukuoka

(December 3-5, 2008)

ABSTRACT

Here, a new codon-based model for amino acid substitutions in homologous proteins is developed and the significance of the simultaneous changes of multiple nucleotides in a codon is shown contrary to a common concept that amino acid substitutions proceed in a stepwise manner. In the present codon-base model, mutational trends at the nucleotide level, the effects of a genetic code, and selective constrains on amino acids are separated from each other. The selective constraints on amino acids are primarily characteristics of amino acids and secondarily proteins but not those of organelles and species. The mutational process of individual nucleotides is modeled as a general reversible Markov process by assuming the substitution rates of codons to be proportional to equilibrium codon frequencies. Codon replacements that require multiple base changes are assumed to occur with the same order of time as single base changes. All codon replacements are assumed to be lethal or neutral, and the ratio of neutral substitutions is assumed to depend on the type of amino acid pair. This codon-based model is examined by maximizing its likelihoods to the amino acid substitution matrices, the JTT and WAG for nuclear proteins, the mtREV estimated for vertebrate mitochondria encoded proteins, and the cpREV for chloroplast encoded proteins.

With and without the simultaneous changes of multiple nucleotides, 75 parameters for the ratios of neutral substitutions for all single step amino acid pairs and 6 parameters for 6 groups of multi-step amino acid pairs are estimated by maximizing the likelihood to the JTT. The Akaike information criterion (AIC) is significantly improved by assuming the simultaneous changes of multiple nucleotides. The maximum likelihood (ML) estimators of the 75 parameters for all single step amino acid pairs more correlate with their physico-chemical estimates in the model with the assumption than that without the assumption of the simultaneous changes of multiple nucleotides. Also, in the model without the assumption, the ML estimators of some other parameters take an unreasonable value, indicating that a model with appropriate assumptions must be employed to obtain biological knowledge from ML estimators.

This set of the ML estimators of the ratios of neutral substitutions for the JTT can almost perfectly reproduce the JTT, and significantly better fit the cpREV, mtREV and WAG than other estimates based on physico-chemical characteristics of amino acids. Also, other parameters at the DNA level can be adjusted to increase likelihoods in phylogenetic analyses than the JTT. Thus, the present codon-based model with these better estimates of the parameters would be useful as a simple evolutionary model for phylogenetic estimation, which allows us to analyze not only amino acid sequences but also codon sequences without any loss of information, and also useful to generate log-odds for codon substitutions for the evolution of any DNA sequence encoding proteins.

1. INTRODUCTION

Purpose and distinctive features of the present study:

- To develop a codon-based Markov model for amino acid substitutions in protein evolution.
  - Mutational trends at the nucleotide level, and selection at the amino acid level, and the genetic code connecting their two levels are taken into account.
  - Unlike previous models, codon replacements by the simultaneous changes of multiple nucleotides are taken into account.
- To confirm the significance of the simultaneous changes of multiple nucleotides in amino acid substitutions.
- To estimate the ratio of neutral substitutions for each amino acid pair by maximizing the likelihood to the JTT substitution matrix.
- To confirm that the present model with the ML estimators of the ratios of neutral substitutions better fit the observed substitution matrices such as the cpREV, mtREV and WAG than that with their physico-chemical estimates and even the JTT.
- To confirm that the whole evolutionary process of amino acid substitutions is encoded in a transition matrix evaluated by the present Markov model.

2. METHODS

A stationary Markov model for codon substitutions

Transition matrix over time t:	$S(t) = \exp(Rt)$	with $f_\mu R_{\mu\nu} = f_\nu R_{\nu\mu}$
Substitution rate matrix:	$R_{\mu\nu} = \text{const}\, M_{\mu\nu} \frac{f_\nu}{f_\nu^{\text{mut}}} e^{w_{\mu\nu}}$	for $\mu \neq \nu$, normalized to $\sum_\mu f_\mu R_{\mu\mu} = -1$
Mutation rate matrix:	$M_{\mu\nu} = \prod_{i=1}^{3} [\delta_{\mu_i\nu_i} + (1 - \delta_{\mu_i\nu_i})(r_i)_{\mu_i\nu_i} f_{i,\nu_i}^{\text{mut}}]$	for $\mu \neq \nu$
Selective constraints:	$e^{w_{\mu\nu}} = \sum_a \sum_b C_{\mu a} C_{\nu b} e^{w_{ab}}$	for $a, b \in$ amino acids

where

f_μ	Equilibrium frequency of codon μ
f_μ^{mut}	Equilibrium frequency of codon μ for the M; $f_{\mu=(\mu_1,\mu_2,\mu_3)}^{\text{mut}} = f_{1,\mu_1}^{\text{mut}} f_{2,\mu_2}^{\text{mut}} f_{3,\mu_3}^{\text{mut}}$
$f_{\mu_i}^{\text{mut}}$	Equilibrium frequency of nucleotide μ_i at codon position i for the M
$w_{ab} = w_{ba}$	Selective constraint against substitutions between amino acids a and b; $w_{aa} = 0$ and $w_{ab} < 0$ for $a \neq b$
$C_{\mu a}$	Genetic code table; $C_{\mu a} = 1$ if μ is a codon for amino acid a, otherwise 0
$r_{\mu_i\nu_i}$	Relative mutation rate between nucleotides μ_i and ν_i at codon position i; $r_{\mu_i\nu_i} = r_{\nu_i\mu_i}$
	$\mu_i \in \{t, c, a, g\}$, $a, b \in \{\text{amino acids}\}$

For simplicity, $r_{\mu_i\nu_i}$ and $f_{\mu_i}^{\text{mut}}$ are assume to be independent of i.

Likelihood to an observed substitution matrix

Log-likelihood: $\ell(\boldsymbol{\theta}) = N \sum_a \sum_b \hat{f}_a S_{ab}^{\text{obs}} \log(f_a \langle S \rangle(\tau, \sigma)_{ab})$

Kullback-Leibler Information: $\hat{I}_{\text{KL}}(\boldsymbol{\theta}) = -\ell(\boldsymbol{\theta}) + \sum_a \sum_b \hat{f}_a S_{ab}^{\text{obs}} \log(\hat{f}_a S_{ab}^{\text{obs}})$

Mean of $S(t)$ over t: $\langle S \rangle(\tau, \sigma) = \int_0^\infty S(t)\Gamma(t; \tau, \sigma)dt = [(I - \sigma R)^{-1}]^\tau$

Estimates of parameters:

Maximum log-likelihood: $\ell(\hat{\boldsymbol{\theta}})) \equiv \max_{\boldsymbol{\theta}} \ell(\boldsymbol{\theta})$, $\hat{I}_{\text{KL}}(\hat{\boldsymbol{\theta}}) \equiv \min_{\boldsymbol{\theta}} \hat{I}_{\text{KL}}(\boldsymbol{\theta})$

Codon frequencies: $C_{\mu a}\hat{f}_\mu = \hat{f}_a C_{\mu a} f_\mu^{\text{usage}} \;/\; \sum_\nu C_{\nu a} f_\nu^{\text{usage}}$, $f^{\text{usage}}_{\mu=(\mu_1,\mu_2,\mu_3)} = f^{\text{usage}}_{\mu_1} f^{\text{usage}}_{\mu_2} f^{\text{usage}}_{\mu_3}$

Observed substitution matrix: $A_{ab} = N \hat{f}_a S_{ab}^{\text{obs}}$

Shape parameter $\hat{\tau}$ of Γ: $\sum_a \hat{f}_a \langle S(\hat{\tau}, \sigma) \rangle_{aa} = \sum_a \hat{f}_a S_{ab}^{\text{obs}}$

Evaluation of model:

Akaike Information Criterion: $\text{AIC} \equiv -2\ell(\hat{\boldsymbol{\theta}}) + 2 \cdot (\text{number of adjustable parameters})$

Log-odds:

$$\text{log-}O(\langle S \rangle(t))_{ab} \equiv \frac{10}{\log 10} \log \frac{\langle S \rangle(t)_{ab}}{f_b} \qquad \mu_i \in \{t, c, a, g\} \text{ ,} \quad a, b \in \{\text{amino acids}\}$$

Observed transition matrices used for model fitting

JTT matrix: compiled from closely related proteins by Jones et al. (1992)

cpREV matrix: estimated from chloroplast proteins (Adachi et al., 2000)

mtREV matrix: estimated from vertebrate mitochondrial proteins (Adachi & Hasegawa, 1996)
by maximizing the likelihood to a given phylogenetic tree.

WAG matrix: estimated from proteins encoded in nuclear DNA (Whelan &Goldman, 2001)
by maximizing the likelihood to given phylogenetic trees and branch lengths.

Parameters

- 190 parameters: w_{ab} for selective constraints on amino acids
- 1 parameter for the simultaneous change of multiple nucleotides; $r_{[tc][ag]}$
- 5 parameters: relative rates; $r_{tc|ag}/r_{[tc][ag]}$, $r_{ag}/r_{tc|ag}$, $r_{ta}/r_{[tc][ag]}$, $r_{tg}/r_{[tc][ag]}$, $r_{ca}/r_{[tc][ag]}$
- 6 parameters: nucleotide frequencies $f^{\text{mut}}_{[atcg]}$ and codon usage $f^{\text{usage}}_{[atcg]}$
- 1 parameter: the scale parameter σ of a Γ distribution

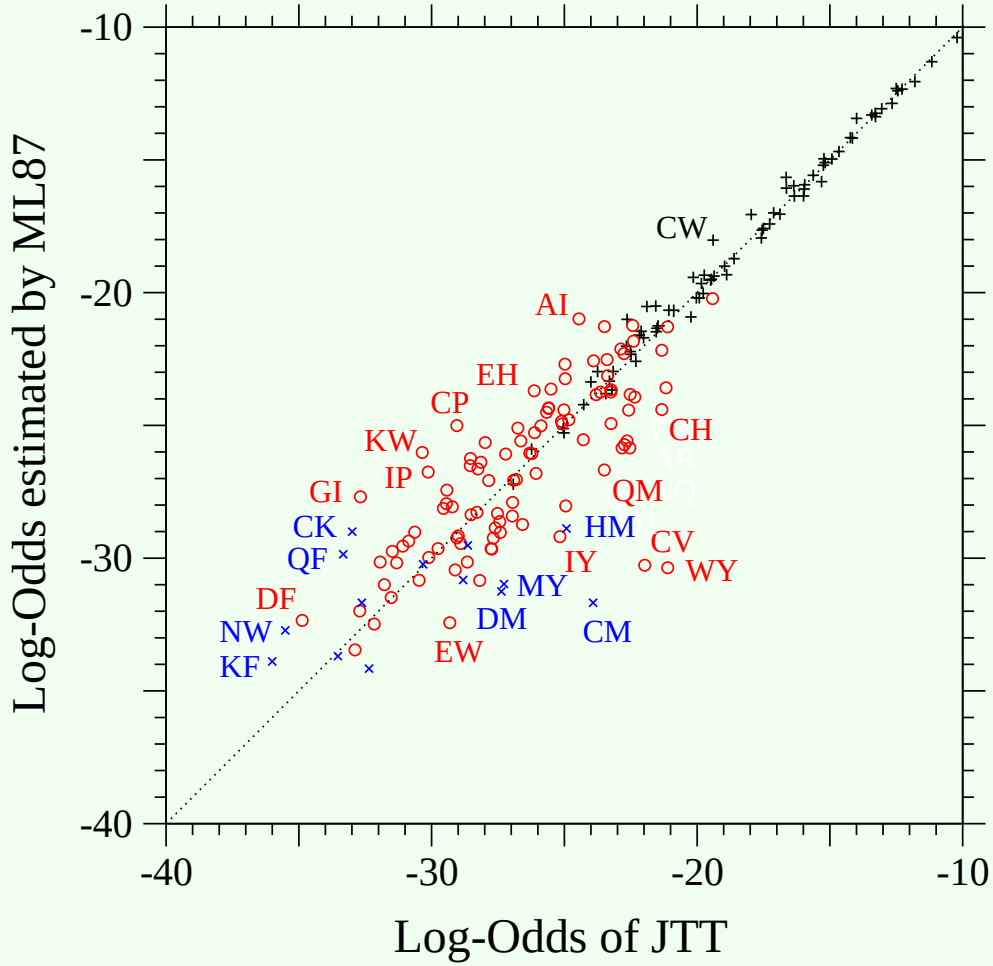
20 parameters taken to be equal to their observed values:

- 1 parameter: the shape parameter τ of a Γ distribution
- 19 parameters: amino acid frequencies f_a

3. RESULTS

Log-odds of the ML-87 model fitted to the JTT 1-PAM

All 75 $\hat{w}_{ab}$ for single-step amino acid pairs are optimized with no multiple base changes, $\hat{r}_{[tc][ag]} \to 0$.



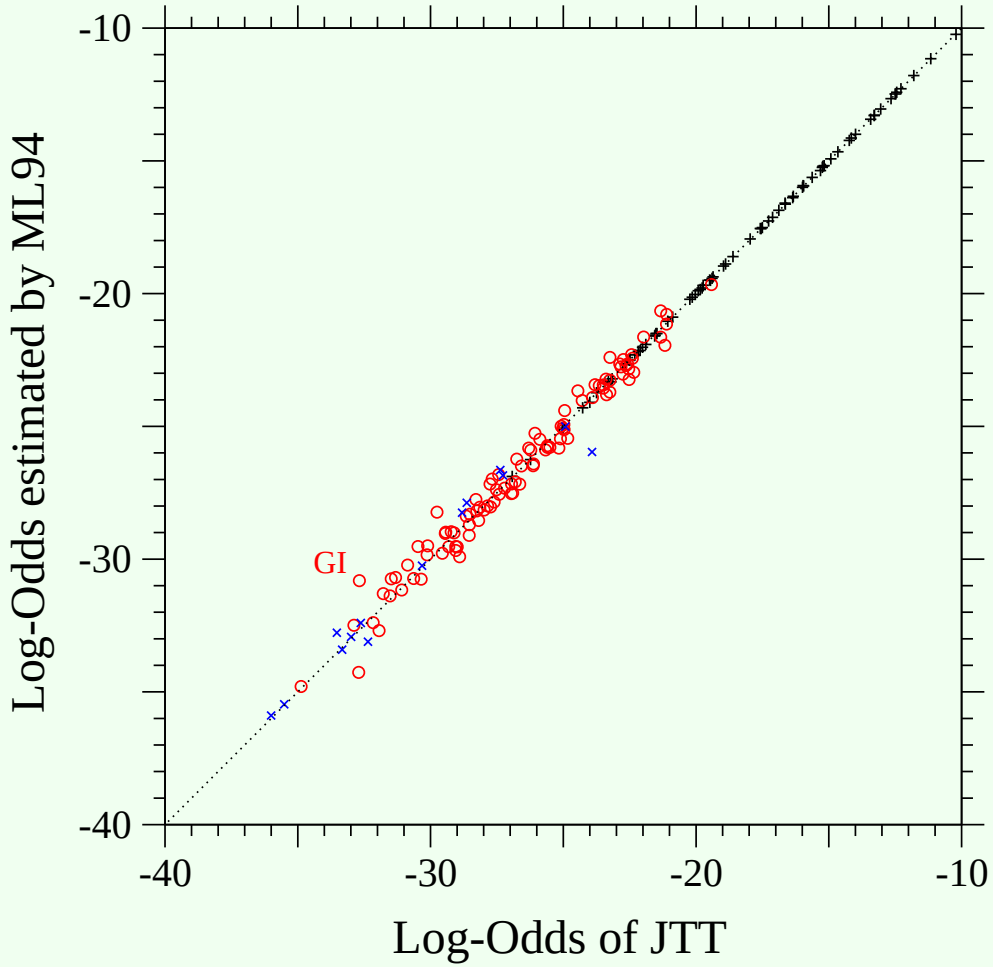
AIC = 35351354.9

+ one nucleotide change o two nucleotide change x three nucleotide change

	ML–87		ML–94	
$\hat{r}_{[tc][ag]}$	$(\to 0)$		0.640	In ML-87, substitutions occur only between one step amino acid pairs.
$\hat{r}_{tc\|ag}/\hat{r}_{[tc][ag]}$	0.303		1.57	
$\hat{r}_{ag}/\hat{r}_{tc\|ag}$	1.58		1.15	
$\hat{r}_{ta}/\hat{r}_{[tc][ag]}$	0.0688		0.725	
$\hat{r}_{tg}/\hat{r}_{[tc][ag]}$	3.04		0.940	
$\hat{r}_{ca}/\hat{r}_{[tc][ag]}$	0.583		1.19	
$\hat{f}^{\mathsf{mut}}_{t+a}$	0.514		0.456	
$\hat{f}^{\mathsf{mut}}_{t}/\hat{f}^{\mathsf{mut}}_{t+a}$	0.324		0.502	
$\hat{f}^{\mathsf{mut}}_{c}/\hat{f}^{\mathsf{mut}}_{c+g}$	0.534		0.436	
$\hat{f}^{\mathsf{usage}}_{t+a}$	0.110		0.475	
$\hat{f}^{\mathsf{usage}}_{t}/\hat{f}^{\mathsf{usage}}_{t+a}$	0.514		0.489	
$\hat{f}^{\mathsf{usage}}_{c}/\hat{f}^{\mathsf{usage}}_{c+g}$	0.254		0.540	
$\hat{\sigma}$	7.29		0.743	
$\hat{\tau}\hat{\sigma}$	0.0886		0.0243	
#parameters	107		114	
$\hat{I}_{KL}(\hat{\boldsymbol{\theta}})^{\dagger}$	0.00016600		0.00000624	
AIC‡	35351314.8	$>$	35349437.7	indicates that the ML-94 better fits the JTT than the ML-87.
Ratio of substitution rates				
the total base/codon	1.30		1.35	
transition/transversion	0.583		1.06	
nonsynonymous/synonymous	1.33		1.37	
Ratio of substitution rates for $\sigma \to 0$				
total base/codon	1.0		1.22	
transition/transversion	0.310		1.19	
nonsynonymous/synonymous	0.283		1.05	

Log-odds of the ML-94 model fitted to the JTT 1-PAM

All 75 $\hat{w}_{ab}$ for single-step amino acid pairs and 6 categorized $\hat{w}_{ab}$ for multi-step are optimized.

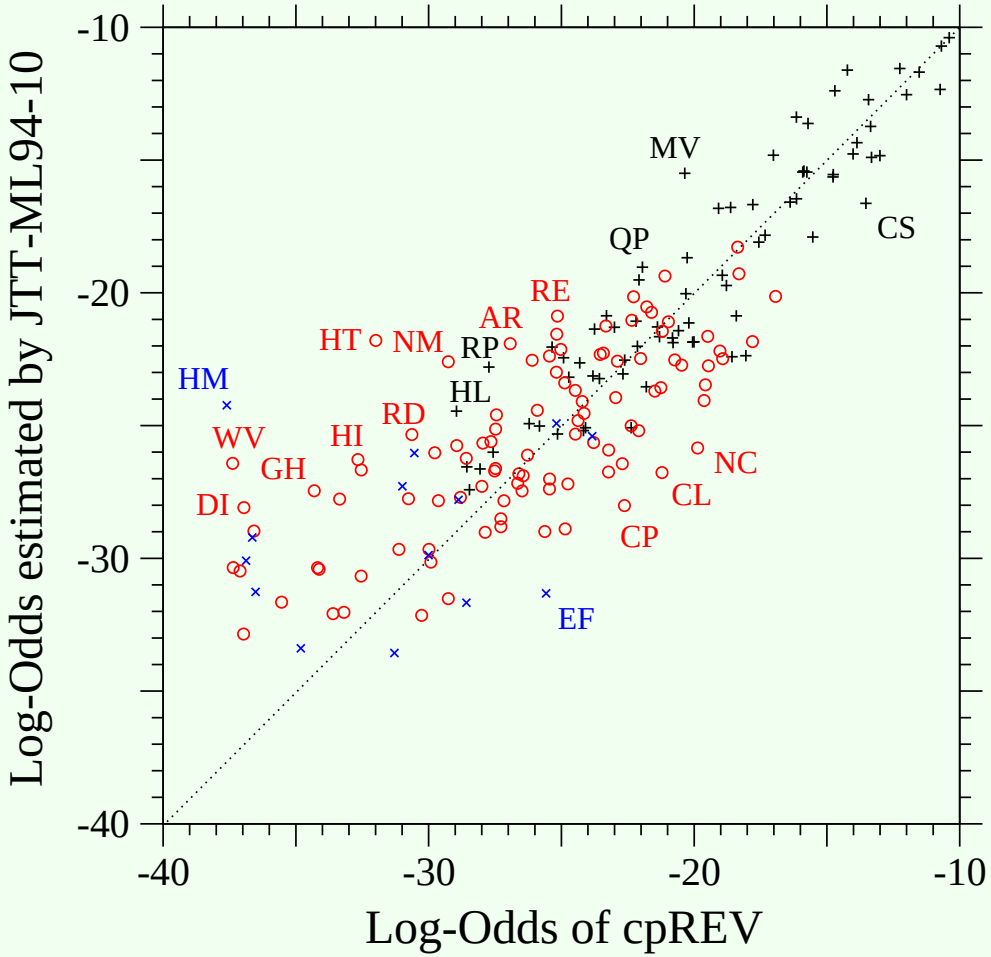


AIC = 35349437.7

\+ one nucleotide change o two nucleotide change x three nucleotide change

Log-odds of the JTT-ML94-10 model fitted to the cpREV 1-PAM

$\hat{w}^{\text{JTT-ML94}}$ of the ML-94 fitted to the JTT is used; $w_{ab} = \beta \hat{w}_{ab}^{\text{JTT-ML94}}$.



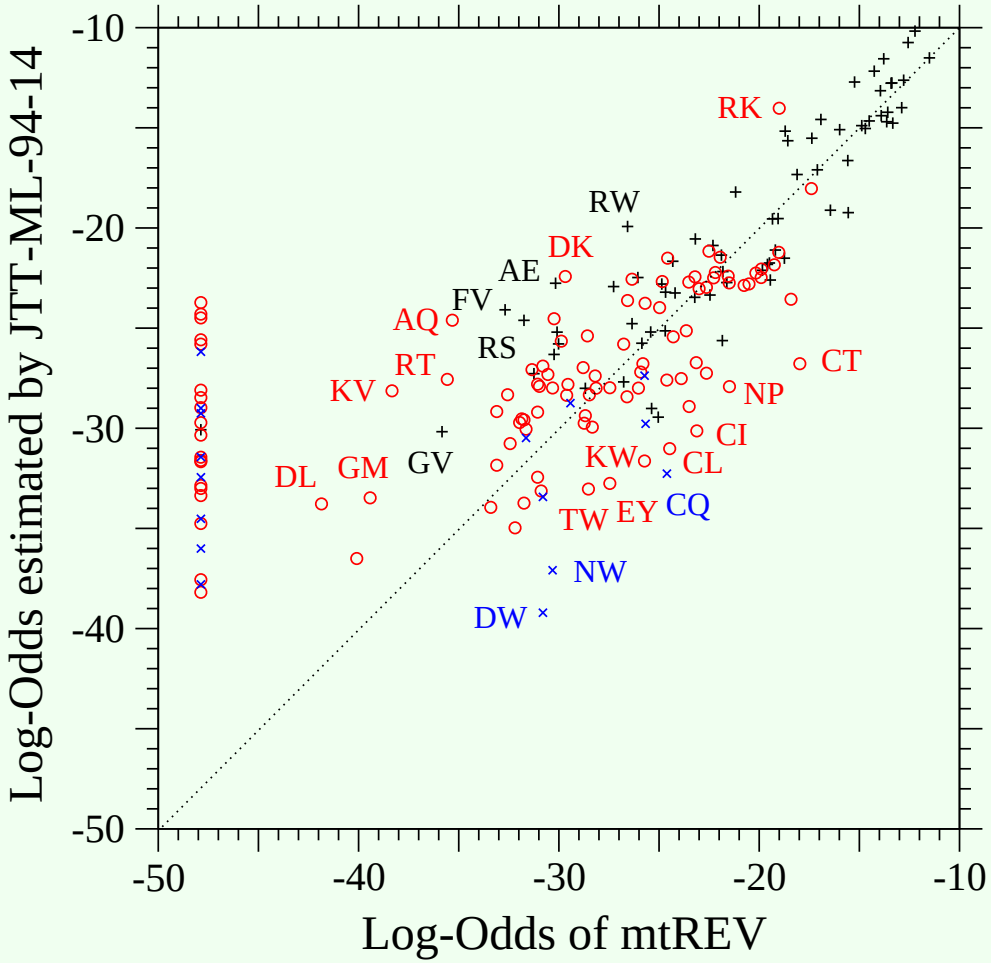
\+ one nucleotide change o two nucleotide change x three nucleotide change

AIC indicates that the JTT-ML94-10 significantly better fits the cpREV than the JTT-ML94-0 (=JTT-F).

	JTT-ML94-⋆							
	0	1	4	8	10	13	14	EI-14
$1/\hat{\beta}$	(1.0)	(1.0)	0.968	0.998	0.865	0.801	0.899	1.87
$\hat{r}_{[tc][ag]}$	(0.640)	(0.640)	0.543	0.546	0.834	$(\to 0)$	1.04	0.872
$\hat{r}_{tc\|ag}/\hat{r}_{[tc][ag]}$	(1.57)	(1.57)	1.59	1.55	1.54	1.50	1.52	1.78
$\hat{r}_{ag}/\hat{r}_{tc\|ag}$	(1.15)	(1.15)	(1.15)	1.30	(1.15)	1.32	1.20	0.996
$\hat{r}_{ta}/\hat{r}_{[tc][ag]}$	(0.725)	(0.725)	(0.725)	0.666	(0.725)	0.585	0.746	1.32
$\hat{r}_{tg}/\hat{r}_{[tc][ag]}$	(0.940)	(0.940)	(0.940)	1.03	(0.940)	1.39	1.37	1.49
$\hat{r}_{ca}/\hat{r}_{[tc][ag]}$	(1.19)	(1.19)	(1.19)	1.46	(1.19)	1.29	1.18	0.760
$\hat{f}^{\mathrm{mut}}_{t+a}$	(0.456)	(0.456)	(0.456)	(0.456)	0.272	0.277	0.268	0.390
$\hat{f}^{\mathrm{mut}}_{t}/\hat{f}^{\mathrm{mut}}_{t+a}$	(0.502)	(0.502)	(0.502)	(0.502)	0.594	0.584	0.570	0.621
$\hat{f}^{\mathrm{mut}}_{c}/\hat{f}^{\mathrm{mut}}_{c+g}$	(0.436)	(0.436)	(0.436)	(0.436)	0.415	0.445	0.465	0.507
$\hat{f}^{\mathrm{usage}}_{t+a}$	(0.475)	(0.475)	(0.475)	(0.475)	0.515	0.582	0.314	0.653
$\hat{f}^{\mathrm{usage}}_{t}/\hat{f}^{\mathrm{usage}}_{t+a}$	(0.489)	(0.489)	(0.489)	(0.489)	0.442	0.469	0.0237	0.0261
$\hat{f}^{\mathrm{usage}}_{c}/\hat{f}^{\mathrm{usage}}_{c+g}$	(0.540)	(0.540)	(0.540)	(0.540)	0.785	0.660	0.990	0.966
$\hat{\sigma}$	(0.743)	1.62	2.04	1.96	2.27	8.79	1.20	0.811
$\hat{\tau}\hat{\sigma}$	0.0249	0.0295	0.0329	0.0319	0.0347	0.0741	0.0216	0.0166
#parameters	20	21	24	28	30	33	34	34
$\hat{I}_{KL}(\hat{\boldsymbol{\theta}})^{\dagger}$	0.00126338	0.00116929	0.00116609	0.00107810	0.00066691	0.00079838	0.00055685	0.00139995
AIC‡	884004.7	883978.6	883983.6	883965.3	883846.5	883891.8	883821.6	884073.5
Ratio of substitution rates								
the total base/codon	1.35	1.43	1.42	1.42	1.44	1.42	1.41	1.36
transition/transversion	1.06	0.984	0.986	0.959	0.963	0.989	0.535	0.731
non-/synonymous	1.31	1.56	1.51	1.58	1.78	1.71	4.34	8.11
For $\sigma \to 0$								
the total base/codon	1.22	1.22	1.18	1.19	1.18	1.0	1.22	1.22
transition/transversion	1.17	1.17	1.20	1.17	1.30	1.73	0.600	0.849
non-/synonymous	1.00	1.00	0.904	0.974	0.996	0.700	2.75	6.87

Log-odds of the JTT-ML94-14 model fitted to the mtREV 1-PAM

$\hat{w}^{\text{JTT-ML94}}$ of the ML-94 fitted to the JTT is used; $w_{ab} = \beta \hat{w}_{ab}^{\text{JTT-ML94}}$.



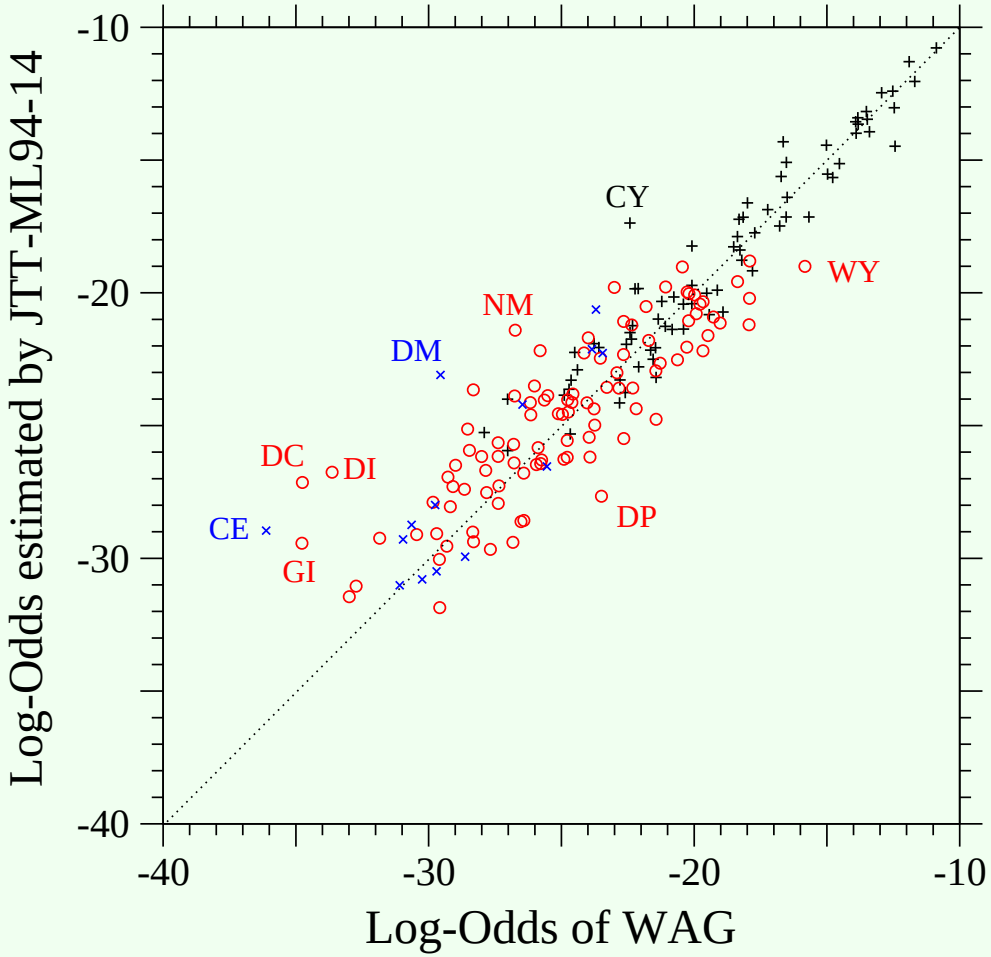
\+ one nucleotide change o two nucleotide change x three nucleotide change

AIC indicates that the JTT-ML94-14 significantly better fits the mtREV than the JTT-ML94-0 (=JTT-F).

	JTT-ML94-⋆							
	0	1	4	8	10	13	14	EI-14
$1/\hat{\beta}$	(1.0)	(1.0)	0.795	0.745	0.658	0.631	0.661	2.01
$\hat{r}_{[tc][ag]}$	(0.640)	(0.640)	0.420	0.356	0.541	$(\to 0)$	0.470	0.875
$\hat{r}_{tc\|ag}/\hat{r}_{[tc][ag]}$	(1.57)	(1.57)	1.94	2.01	2.17	2.40	2.26	3.58
$\hat{r}_{ag}/\hat{r}_{tc\|ag}$	(1.15)	(1.15)	(1.15)	0.984	(1.15)	1.14	1.08	0.828
$\hat{r}_{ta}/\hat{r}_{[tc][ag]}$	(0.725)	(0.725)	(0.725)	0.718	(0.725)	0.789	0.825	1.50
$\hat{r}_{tg}/\hat{r}_{[tc][ag]}$	(0.940)	(0.940)	(0.940)	0.698	(0.940)	0.761	0.817	0.354
$\hat{r}_{ca}/\hat{r}_{[tc][ag]}$	(1.19)	(1.19)	(1.19)	1.85	(1.19)	1.69	1.55	0.706
$\hat{f}^{\mathrm{mut}}_{t+a}$	(0.456)	(0.456)	(0.456)	(0.456)	0.223	0.251	0.254	0.303
$\hat{f}^{\mathrm{mut}}_{t}/\hat{f}^{\mathrm{mut}}_{t+a}$	(0.502)	(0.502)	(0.502)	(0.502)	0.588	0.601	0.593	0.637
$\hat{f}^{\mathrm{mut}}_{c}/\hat{f}^{\mathrm{mut}}_{c+g}$	(0.436)	(0.436)	(0.436)	(0.436)	0.351	0.356	0.362	0.285
$\hat{f}^{\mathrm{usage}}_{t+a}$	(0.475)	(0.475)	(0.475)	(0.475)	0.458	0.533	0.474	0.954
$\hat{f}^{\mathrm{usage}}_{t}/\hat{f}^{\mathrm{usage}}_{t+a}$	(0.489)	(0.489)	(0.489)	(0.489)	0.378	0.433	0.378	0.678
$\hat{f}^{\mathrm{usage}}_{c}/\hat{f}^{\mathrm{usage}}_{c+g}$	(0.540)	(0.540)	(0.540)	(0.540)	0.800	0.756	0.804	0.187
$\hat{\sigma}$	(0.743)	1.03	2.51	2.73	3.90	8.02	3.47	0.587
$\hat{\tau}\hat{\sigma}$	0.0276	0.0293	0.0465	0.0512	0.0610	0.105	0.0579	0.0156
#parameters	20	21	24	28	30	33	34	34
$\hat{I}_{KL}(\hat{\boldsymbol{\theta}})^{\dagger}$	0.00141210	0.00139587	0.00128306	0.00111284	0.00078145	0.00084340	0.00073257	0.00198631
AIC‡	785822.9	785820.5	785795.4	785756.6	785669.3	785692.4	785663.9	786009.0
Ratio of substitution rates								
the total base/codon	1.33	1.36	1.35	1.34	1.34	1.33	1.34	1.39
transition/transversion	1.08	1.05	1.24	1.30	1.21	1.30	1.20	0.725
non-/synonymous	1.09	1.17	0.958	0.878	1.06	0.938	1.02	7.68
For $\sigma \to 0$								
the total base/codon	1.20	1.20	1.12	1.11	1.09	1.0	1.09	1.27
transition/transversion	1.21	1.21	1.68	1.80	2.13	2.67	1.96	0.798
non-/synonymous	0.798	0.798	0.500	0.431	0.416	0.285	0.421	6.11

Log-odds of the JTT-ML94-14 model fitted to the WAG 1-PAM

$\hat{w}^{\text{JTT-ML94}}$ of the ML-94 fitted to the JTT is used; $w_{ab} = \beta \hat{w}_{ab}^{\text{JTT-ML94}}$.

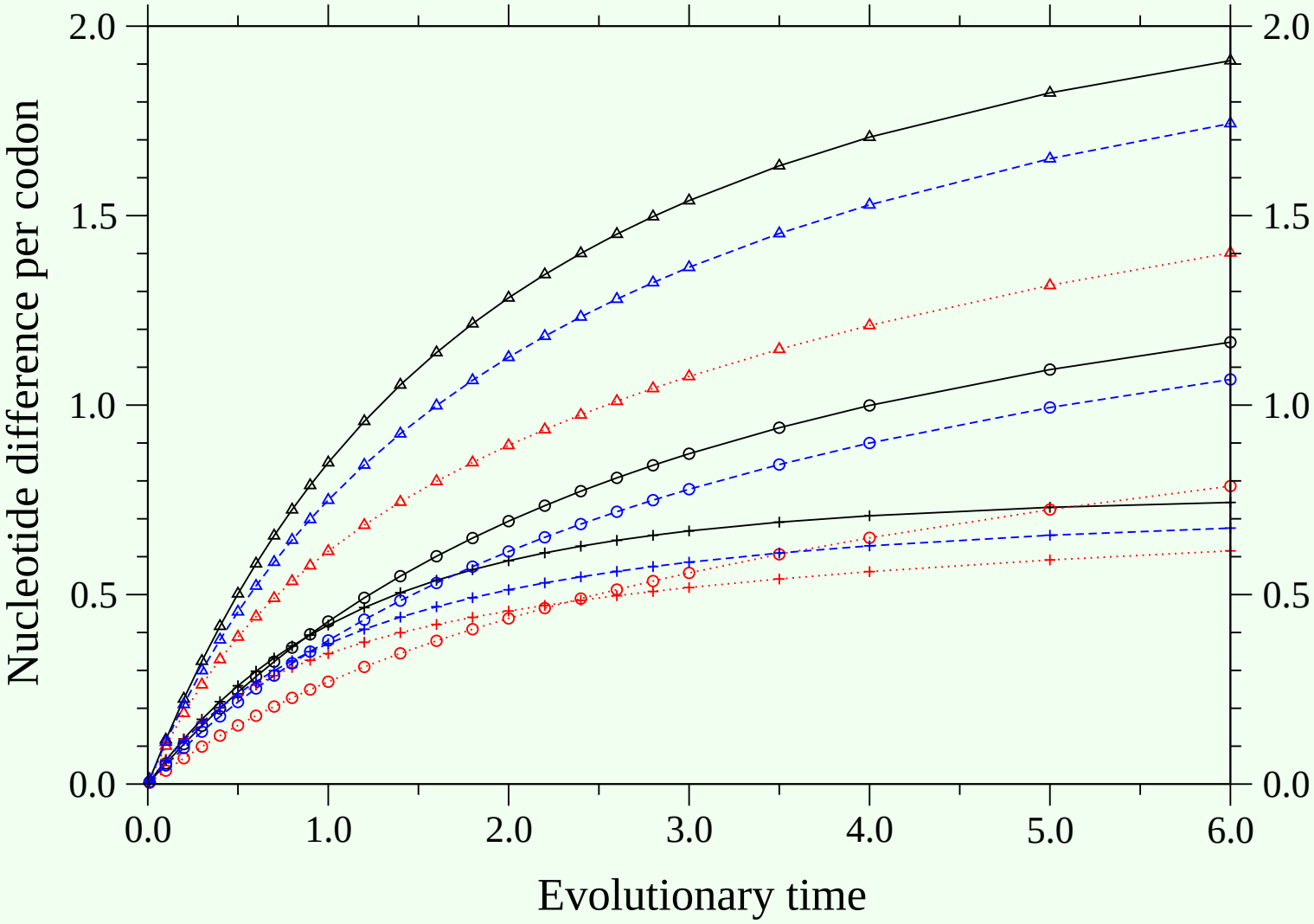


\+ one nucleotide change o two nucleotide change x three nucleotide change

AIC indicates that the JTT-ML94-14 significantly better fits the WAG than the JTT-ML94-0 (=JTT-F).

	JTT-ML94-⋆							
	0	1	4	8	10	13	14	EI-14
$1/\hat{\beta}$	(1.0)	(1.0)	1.04	1.04	0.988	0.763	0.952	1.92
$\hat{r}_{[tc][ag]}$	(0.640)	(0.640)	0.967	0.957	1.30	$(\to 0)$	1.17	1.07
$\hat{r}_{tc\|ag}/\hat{r}_{[tc][ag]}$	(1.57)	(1.57)	1.22	1.20	1.26	1.13	1.26	1.39
$\hat{r}_{ag}/\hat{r}_{tc\|ag}$	(1.15)	(1.15)	(1.15)	1.29	(1.15)	1.35	1.30	1.04
$\hat{r}_{ta}/\hat{r}_{[tc][ag]}$	(0.725)	(0.725)	(0.725)	0.776	(0.725)	0.699	0.813	1.35
$\hat{r}_{tg}/\hat{r}_{[tc][ag]}$	(0.940)	(0.940)	(0.940)	0.801	(0.940)	0.705	0.864	1.22
$\hat{r}_{ca}/\hat{r}_{[tc][ag]}$	(1.19)	(1.19)	(1.19)	1.38	(1.19)	1.71	1.34	0.799
$\hat{f}^{\mathrm{mut}}_{t+a}$	(0.456)	(0.456)	(0.456)	(0.456)	0.321	0.376	0.339	0.456
$\hat{f}^{\mathrm{mut}}_{t}/\hat{f}^{\mathrm{mut}}_{t+a}$	(0.502)	(0.502)	(0.502)	(0.502)	0.530	0.554	0.550	0.557
$\hat{f}^{\mathrm{mut}}_{c}/\hat{f}^{\mathrm{mut}}_{c+g}$	(0.436)	(0.436)	(0.436)	(0.436)	0.471	0.445	0.482	0.525
$\hat{f}^{\mathrm{usage}}_{t+a}$	(0.475)	(0.475)	(0.475)	(0.475)	0.366	0.307	0.144	0.692
$\hat{f}^{\mathrm{usage}}_{t}/\hat{f}^{\mathrm{usage}}_{t+a}$	(0.489)	(0.489)	(0.489)	(0.489)	0.198	0.675	0.695	0.0762
$\hat{f}^{\mathrm{usage}}_{c}/\hat{f}^{\mathrm{usage}}_{c+g}$	(0.540)	(0.540)	(0.540)	(0.540)	0.823	0.283	0.315	0.846
$\hat{\sigma}$	(0.743)	2.27	1.27	1.96	1.02	12.0	1.13	0.943
$\hat{\tau}\hat{\sigma}$	0.0245	0.0323	0.0251	0.0319	0.0214	0.0816	0.0221	0.0176
#parameters	20	21	24	28	30	33	34	34
$\hat{I}_{KL}(\hat{\boldsymbol{\theta}})^{\dagger}$	0.00082757	0.00063138	0.00055070	0.00048999	0.00036901	0.00071695	0.00033882	0.00112355
AIC‡	10806243.0	10805532.5	10805245.4	10805033.0	10804597.6	10805867.3	10804495.9	10807346.0
Ratio of substitution rates								
the total base/codon	1.35	1.47	1.49	1.49	1.50	1.46	1.49	1.47
transition/transversion	1.06	0.948	0.781	0.765	0.633	0.734	0.608	0.652
non-/synonymous	1.35	1.76	1.87	1.92	2.70	1.94	2.70	5.52
For $\sigma \to 0$								
the total base/codon	1.22	1.22	1.30	1.30	1.34	1.0	1.33	1.31
transition/transversion	1.18	1.18	0.862	0.850	0.706	1.14	0.691	0.720
non-/synonymous	1.03	1.03	1.29	1.32	1.92	0.649	1.86	4.41

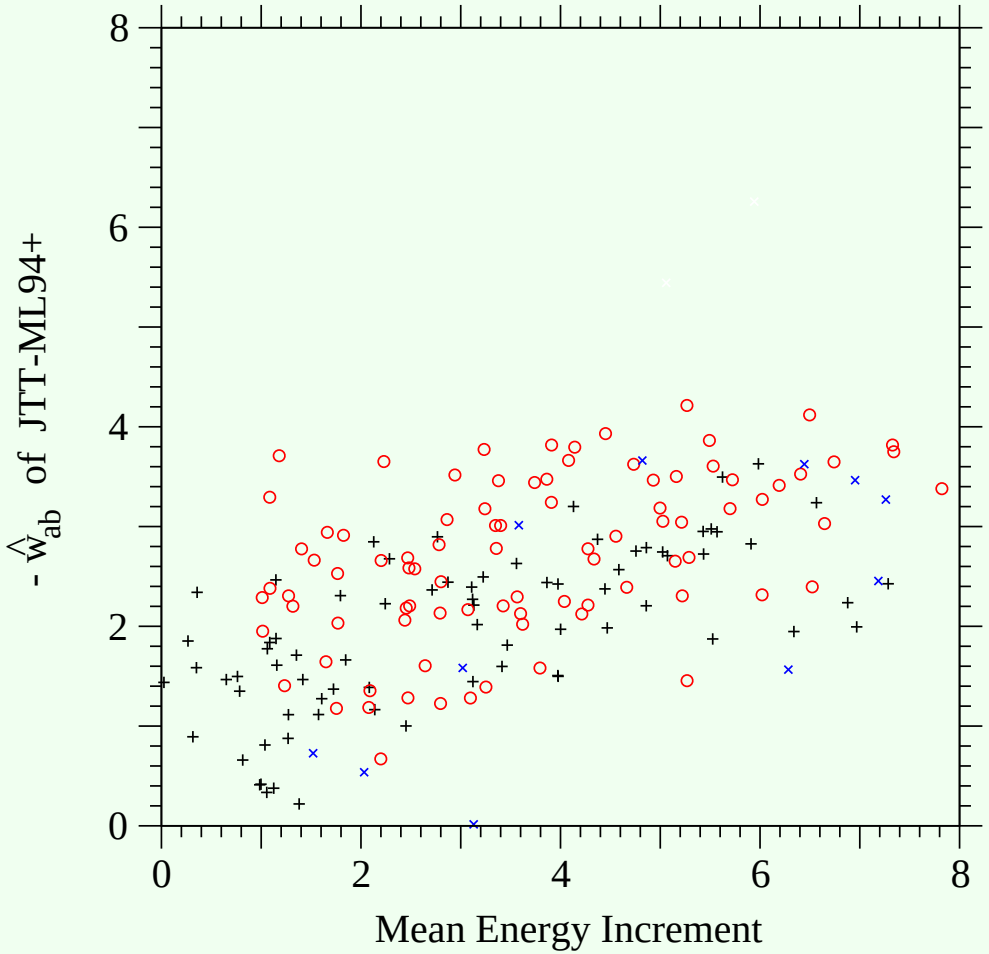
Net changes of transitions and transversions estimated by the present codon-based model



◯ for transversion, + for transition, and △ for the total for the JTT, cpREV, and mtREV

The JTT-ML94 for the JTT, JTT-ML94-10 for the cpREV, and JTT-ML94-13 for the mtREV are used.

$w^{\text{JTT-ML94+}}$ versus mean energy increment due to an amino acid substitution



\+ one nucleotide change o two nucleotide change x three nucleotide change

Correlation coefficients = 0.66, 0.48, 0.70

Temporal changes of eigenvalues and eigenvectors

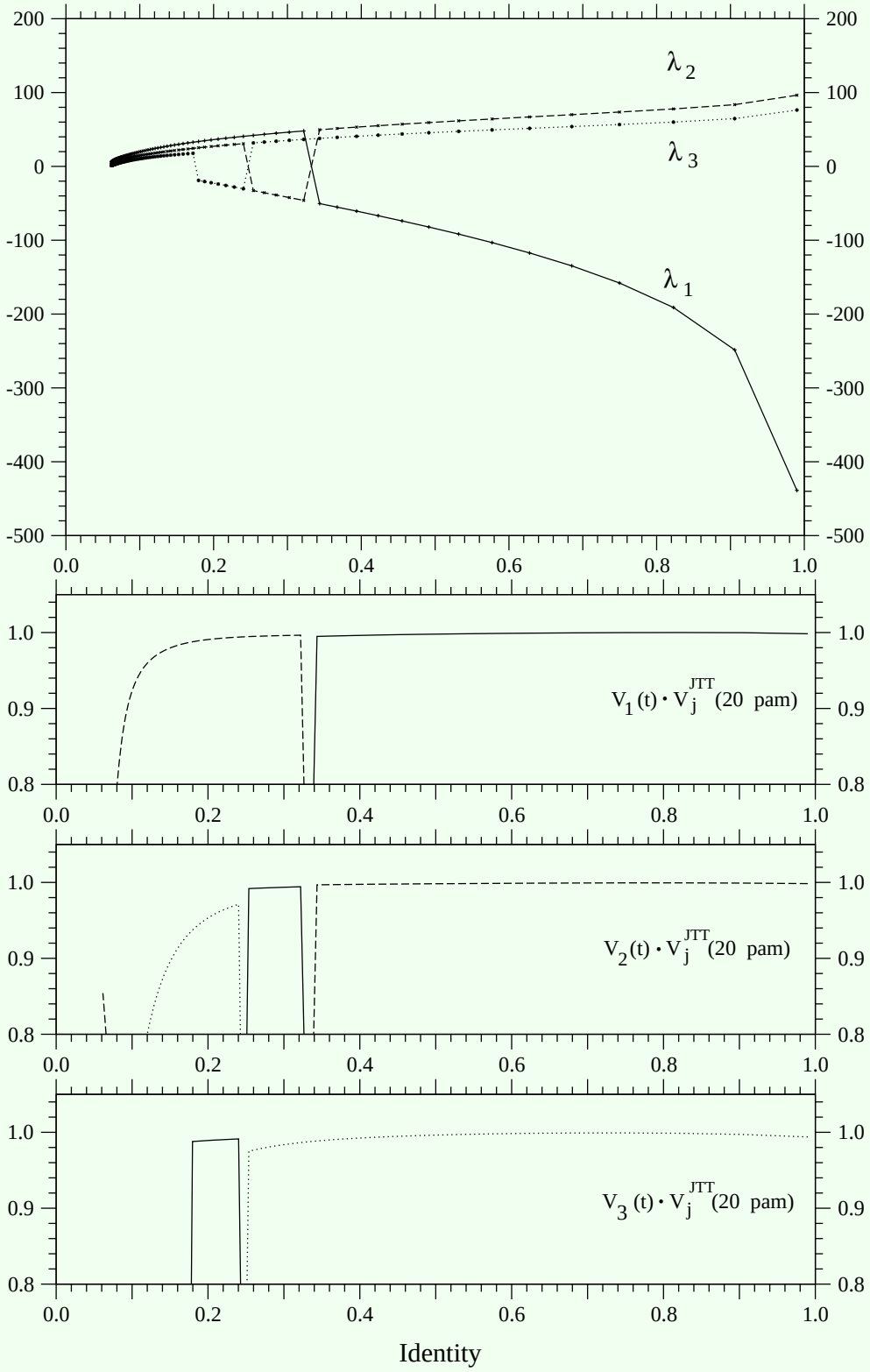
of the log-odds matrix of the ML-94 fitted to the JTT

The whole evolutionary process of amino acid substitutions can be reproduced.

$\longrightarrow$ This fact supports a Markov model for codon substitutions.

The solid, broken, and dotted lines show the 1st, 2nd, and 3rd primary eigenvector.

of the log-odds matrix fitted to the JTT



The solid, broken, and dotted lines show the 1st, 2nd, and 3rd primary eigenvector.

4. CONCLUSION

- The simultaneous changes of multiple nucleotides in codon substitutions are significant, and must be taken into account to model the substitution process of amino acids.
- The JTT amino acid substitution matrix can be almost perfectly reproduced for codon substitutions with the ML estimators (JTT-ML94) of the ratios of neutral substitutions for respective types of amino acid pairs estimated from the JTT.
  - The JTT-ML94 fitted to the JTT can significantly better explain other substitution matrices, such as the cpREV for chloroplast proteins and mtREV for vertebrate mitochondrial proteins, than their physical-chemical estimates.
  - The present codon-based model with the JTT-ML94 can increase likelihoods in phylogenetic analyses with some adjustable parameters at the DNA level than the JTT.
- It has been shown in terms of log-odds matrices that the whole evolutionary process of amino acid substitutions can be reproduced by the transition matrix based on a Markov model.

As a result, the present codon-based model for amino acid substitutions would be useful as a simple evolutionary model for phylogenetic estimation, which allows us to analyze not only amino acid sequences but also codon sequences without any loss of information, and also useful to generate log-odds for codon substitutions for the evolution of any DNA sequences encoding proteins.