Evaluation of Six Methods for Estimating Synonymous and Nonsynonymous Substitution Rates

Zhang Zhang[1,2,3] and Jun Yu[1,2,4*]

[1] *Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China;* [2] *Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 101300, China;* [3] *Graduate School, Chinese Academy of Sciences, Beijing 100039, China;* [4] *James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310007, China.*

Methods for estimating synonymous and nonsynonymous substitution rates among protein-coding sequences adopt different mutation (substitution) models with subtle yet significant differences, which lead to different estimates of evolutionary information. Little attention has been devoted to the comparison of methods for obtaining reliable estimates since the amount of sequence variations within targeted datasets is always unpredictable. To our knowledge, there is little information available in literature about evaluation of these different methods. In this study, we compared six widely used methods and provided with evaluation results using simulated sequences. The results indicate that incorporating sequence features (such as transition/transversion bias and nucleotide/codon frequency bias) into methods could yield better performance. We recommend that conclusions related to or derived from Ka and Ks analyses should not be readily drawn only according to results from one method.

Key words: synonymous substitution, nonsynonymous substitution, Ka/Ks ratio, approximate method, maximum-likelihood method

Introduction

In the field of molecular evolution, one of the powerful tools for understanding the mechanisms of DNA sequence evolution, reconstructing phylogenic trees, and identifying protein-coding exons is to estimate nonsynonymous (amino-acid replacing) and synonymous (silent) substitution rates among protein-coding sequences, termed as Ka and Ks, respectively (*1–5*). Ka reflects nonsynonymous substitutions per nonsynonymous site, and Ks reflects synonymous substitutions per synonymous site. The Ka/Ks ratio (denoted as ω) is widely used as an estimator of selective strength for DNA sequence evolution, with $\omega > 1$ indicating positive selection, $\omega < 1$ indicating purifying (negative) selection, and ω close to 1 indicating neutral mutation.

Over the past two decades, several methods have been developed for Ka and Ks estimations. Although these methods consider different features of sequence evolution, they fall into two classes: approximate methods and maximum-likelihood methods. Approximate methods normally involve three steps to estimate Ka and Ks: Firstly, count the numbers of synonymous (S) and nonsynonymous (N) sites (the sum of S and N is scaled to the length of the sequences compared); Secondly, calculate the numbers of synonymous (S_d) and nonsynonymous (N_d) substitutions (the sum of S_d and N_d equals to the number of substitutions between pairwise sequences); Thirdly, correct for multiple substitutions due to the fact that the observed number of substitutions underestimates the real number of substitutions as sequences diverge over time (*6*). Different from approximate methods, maximum-likelihood methods adopt the probability theory to finish the three steps in one go (*7*). We list the definitions of symbols used in Ka and Ks estimations in Table 1. In addition, these methods can also be classified as nucleotide-based or codon-based methods according to their adopted mutation models. In this study, we focus on six of them: Nei-Gojobori method (NG; ref. *8*), Li-Wu-Luo method (LWL; ref. *9*), Li-Pamilo-Bianchi method (LPB; ref. *10, 11*), Goldman-Yang method (GY; ref. *12*), Yang-Neilsen method (YN; ref. *13*), and modified Yang-

***Corresponding author.**
E-mail: junyu@genomics.org.cn

Table 1 Definitions of Symbols Used in Ka and Ks Estimations

Symbol	Definition
S	Number of synonymous sites
N	Number of nonsynonymous sites
S_d	Number of synonymous substitutions
N_d	Number of nonsynonymous substitutions
Ks	Synonymous substitution rate
Ka	Nonsynonymous substitution rate
ω	Estimator of selective strength, ω = Ka/Ks
t	Divergence time between two sequences, the expected number of nucleotide substitutions per codon, t = (Ks × 3S + Ka × 3N)/(S+N)
κ_R	Ratio of transitional rate between purines to transversional rate
κ_Y	Ratio of transitional rate between pyrimidines to transversional rate
κ	Ratio of transitional rate to transversional rate

Neilsen method (MYN; ref. *14*). Among them, only GY belongs to the maximum-likelihood method.

It should be noted that different methods adopt different mutation models (*12, 15–18*) with subtle yet significant differences, which lead to diverse estimates of evolutionary distance (*19*). Since Ka, Ks, and ω are broadly applied in molecular evolution, it is necessary to evaluate the accuracies of these methods so that evolutionary information among compared sequences can be accurately captured. To our knowledge, few studies have been done on comprehensive evaluation for these six widely used methods. Therefore, we conducted this study to compare and evaluate these methods by computer simulations and empirical data. In addition, we recommend that methods for estimating Ka and Ks should be used cautiously, and conclusions related to or derived from Ka and Ks analyses should not be readily drawn only according to results from one method.

Results

Comparative results

Effects of codon frequency bias and transition/transversion bias

We performed simulations to generate long sequences as consistency analysis (*13*). Since the two ratios of transitional rate between purines (κ_R) and between pyrimidines (κ_Y) to transversional rate often vary from 1.5 to 5, we considered 3.75 as a "typical value". Hence, we can fix one of them to 3.75 and set the other to vary from 1 to 10. We plotted estimates of ω that were calculated with these six methods against κ_R (fixing κ_Y = 3.75) under three different codon frequencies (Figure 1A–I). Similar results can be obtained for fixed κ_R and variable κ_Y (data not shown).

According to the results, codon frequencies have obvious influence on NG, LWL, and LPB, but minor influence on GY, YN, and MYN (Figure 1A–I). Although LWL is more biased than NG in ω estimation, they both have a nearly parallel trend with an increasing κ_R and tend to underestimate ω for most of the parameter combinations examined. These results are in substantial agreement with previous studies (*13, 19*).

Despite the fact that closer results are sometimes estimated for neutral mutation (Figure 1D–F), LPB, which was proposed as a modification of LWL, performs unsteadily as κ_R increases: overestimate ω for purifying selection (Figure 1A–C) and underestimate ω for positive selection (Figure 1G–I). Taking the human codon frequencies as an example, when κ_R = 4 and 10, the estimates of ω given by LPB are 0.316 and 0.345 for ω = 0.3, 0.944 and 0.991 for ω = 1, and 2.380 and 2.406 for ω = 3, respectively. As a whole, LPB has a better performance than NG and LWL.

GY and YN give rise to similar estimates of ω primarily due to the fact that they both take account of major features of DNA sequence evolution (transition/transversion rate bias, nucleotide/codon frequency bias). Ignoring the difference between κ_R and κ_Y, GY and YN produce closer estimates only when $\kappa_R \approx 3.75$. For instance, when κ_R = 4, estimates of ω given by GY and YN under equal codon frequencies are 0.303 and 0.297 for ω = 0.3, 1.010 and 1.012 for ω = 1, and 3.036 and 3.049 for ω = 3, respectively. GY and YN tend to underestimate ω when $\kappa_R < \kappa_Y$ and to overestimate ω when $\kappa_R > \kappa_Y$, and

their biases become more serious as κ_R increases or decreases to extremes (*14*). Compared with NG, LWL, and LPB, GY and YN perform better for most of the parameter combinations tested, which is attributable to the consideration of more evolutionary features.

MYN, a modified YN method, allows for two different ratios of transitional rate between purines (κ_R) and between pyrimidines (κ_Y) to transversional rate as well as nucleotide/codon frequency. It can become equivalent to YN when $\kappa_R = \kappa_Y$ and thus similar results can be observed by the two methods. For example, when $\kappa_R = 4$, estimates of ω by YN and MYN under human codon frequencies are 0.306 and 0.307 for $\omega = 0.3$, 1.025 and 1.026 for $\omega = 1$, and 3.024 and 3.023 for $\omega = 3$, respectively. When $\kappa_R \neq \kappa_Y$, MYN sometimes yields biased estimates, but it represents a better performance for most of the parameter combinations tested.

We also examined Ks estimations and plotted percentage errors of estimated Ks against κ_R (Figure 2A–I; see Materials and Methods). NG and LWL have a tendency to overestimate Ks for most of the parameter settings examined, and the bias of LWL is more serious than that of NG, which is consistent with those found in ω estimations. As to LPB, closer estimates of Ks can be obtained only for neutral mutation (Figure 2D–F). It tends to give rise to negative percentage errors of Ks for purifying selection (Figure 2A–C) and positive percentage errors of Ks for positive selection (Figure 2G–I). These results also agree well with ω estimations.

GY and YN produce similar results of Ks for most of the parameter combinations: closer estimation only when $\kappa_R \approx 3.75$, overestimation when $\kappa_R < \kappa_Y$ (not apparent in Figure 2B), and underestimation when $\kappa_R > \kappa_Y$. In comparison with NG, LWL, and LPB that do not allow for transition/transversion bias or nucleotide/codon frequency bias, GY and YN both perform better in Ks estimations. MYN gives estimates of Ks similar to GY and YN when $\kappa_R \approx \kappa_Y$, which is in agreement with ω estimations. Taking the human codon frequencies as an example, when $\kappa_R = 4$, the percentage errors of estimated Ks calculated with GY, YN, and MYN are −3.957%, −1.991%, and −2.114% for $\omega = 0.3$, −2.184%, −3.770%, and −3.778% for $\omega = 1$, and −0.709%, −2.614%, and −2.558% for $\omega = 3$, respectively. MYN yields biased estimates when $\kappa_R < \kappa_Y$ and closer estimates when $\kappa_R > \kappa_Y$. For instance, when $\kappa_R = 1$, the percentage errors of estimated Ks calculated with GY, YN, and MYN are −3.397%, 0.246%, and −9.156% for $\omega = 0.3$, 1.582%, 1.001%, and −7.544% for $\omega = 1$, and 5.779%, 4.383%, and −4.053% for $\omega = 3$, respectively. Similarly, when $\kappa_R = 10$, those with GY, YN, and MYN are −13.949%, −14.498%, and −7.114% for $\omega = 0.3$, −9.991%, −11.787%, and −5.367% for $\omega = 1$, and −6.386%, −8.213%, and −2.006% for $\omega = 3$, respectively. As a whole, MYN is less biased than other methods for most of the parameter combinations examined.

Estimates of Ka with these six methods were also tested (Figure 3A–I). NG and LWL underestimate Ka, which is consistent with those found in Ks and ω estimations. Compared with LWL, NG gives slightly better estimates of Ka, and they both are more biased than other methods. LPB has a tendency to overestimate Ka for purifying selection [Figure 3A–C; it is not apparent because of slight underestimations of Ka and Ks arising from about 4% loss of sites due to mutations leading to stop codons (*20*)], and to underestimate Ka for neutral mutation and positive selection (Figure 3D–I). GY, YN, and MYN perform similarly and give rise to less bias in Ka estimation. Taking the human codon frequencies as an example, the percentage errors of estimated Ka calculated with GY, YN, and MYN for the expected $\omega = 3$ (Figure 3H) are −6.379%, −7.134%, and −3.466% when $\kappa_R = 1$, −0.749%, −1.851%, and −1.802% when $\kappa_R = 4$, and −1.269%, −2.826%, and −5.638% when $\kappa_R = 10$, respectively. Biases of Ka given by these methods are overall relatively smaller when compared with Ks and ω estimations.

Effects of divergence time

Since the amount of sequence variations reflected in divergence time (t) is always unpredictable within targeted datasets, we performed simulations to examine its effect under the human codon frequencies. Three different combinations of κ_R and κ_Y were tested: κ_R (=1) < κ_Y (=10), κ_R (=10) > κ_Y (=1), and $\kappa_R = \kappa_Y$ (= 3.75). We plotted estimates of ω against t varying from 0.1 to 1 for the expected $\omega = 0.3$, 1, and 3, respectively (Figure 4A–I).

With t increasing, NG and LWL tend to give better estimates of ω for purifying selection and biased estimates for positive selection, whereas t has no obvious influence on them for neutral mutation. LPB overestimates ω for purifying selection and underestimates ω for positive selection, which is consistent with those found above. GY and YN represent a similar trend with t increasing: underestimate ω when

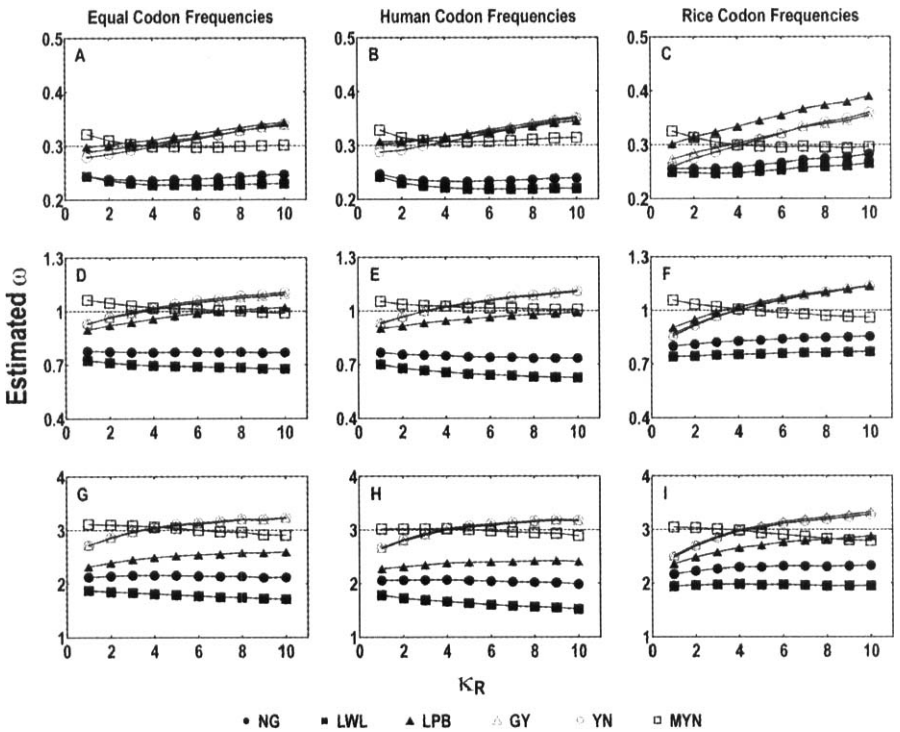

Fig. 1 Estimated ω with six methods when $\kappa_Y = 3.75$, considering κ_R varying from 1 to 10. Three sets of codon frequencies are used: equal (A, D, G), human (B, E, H) calculated from human protein-coding genes, and rice (C, F, I) derived from rice protein-coding genes. $\omega = 0.3$ (A–C), $\omega = 1$ (D–F), and $\omega = 3$ (G–I) are considered as typical values for purifying selection, neutral mutation, and positive selection, respectively.

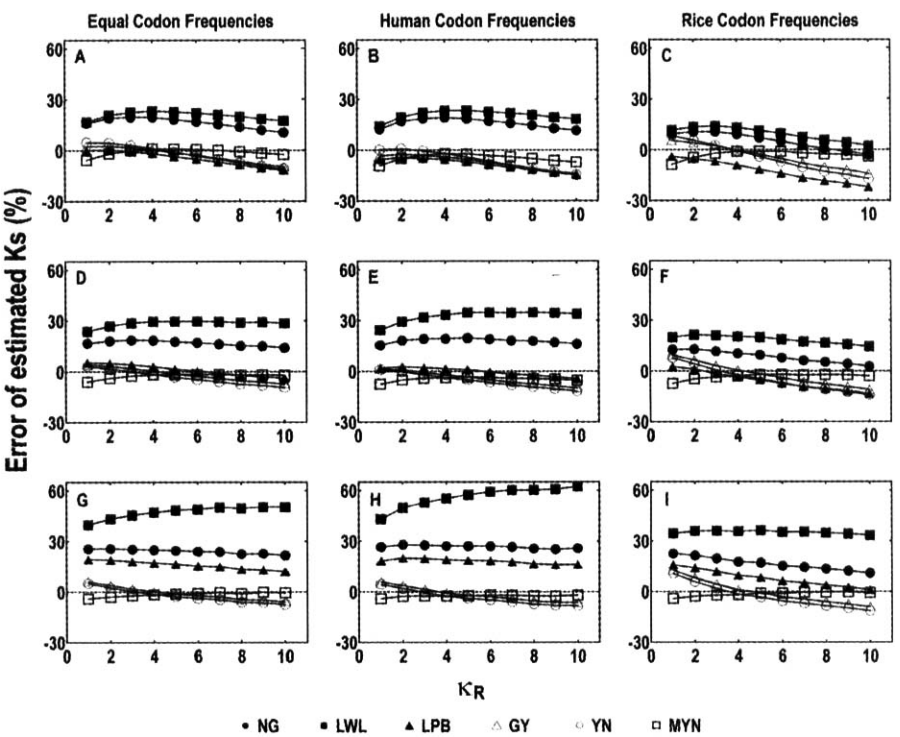

Fig. 2 Percentage errors of estimated Ks with six methods when $\kappa_Y = 3.75$, considering κ_R varying from 1 to 10. Three sets of codon frequencies are used: equal (A, D, G), human (B, E, H) calculated from human protein-coding genes, and rice (C, F, I) derived from rice protein-coding genes. $\omega = 0.3$ (A–C), $\omega = 1$ (D–F), and $\omega = 3$ (G–I) are considered as typical values for purifying selection, neutral mutation, and positive selection, respectively.

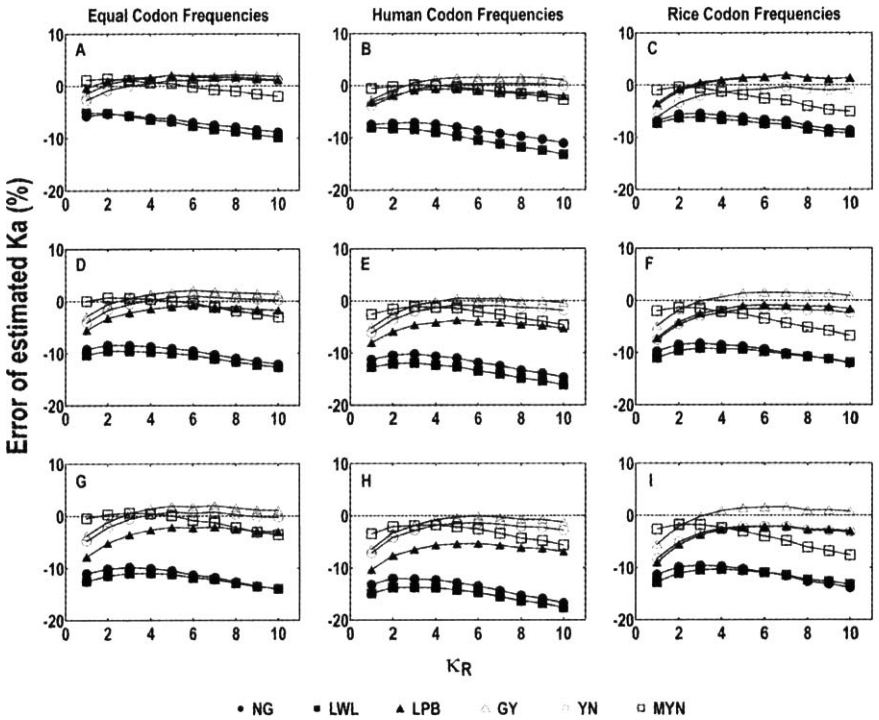

Fig. 3 Percentage errors of estimated Ka with six methods when $\kappa_Y = 3.75$, considering κ_R varying from 1 to 10. Three sets of codon frequencies are used: equal (A, D, G), human (B, E, H) calculated from human protein-coding genes, and rice (C, F, I) derived from rice protein-coding genes. $\omega = 0.3$ (A–C), $\omega = 1$ (D–F), and $\omega = 3$ (G–I) are considered as typical values for purifying selection, neutral mutation, and positive selection, respectively.

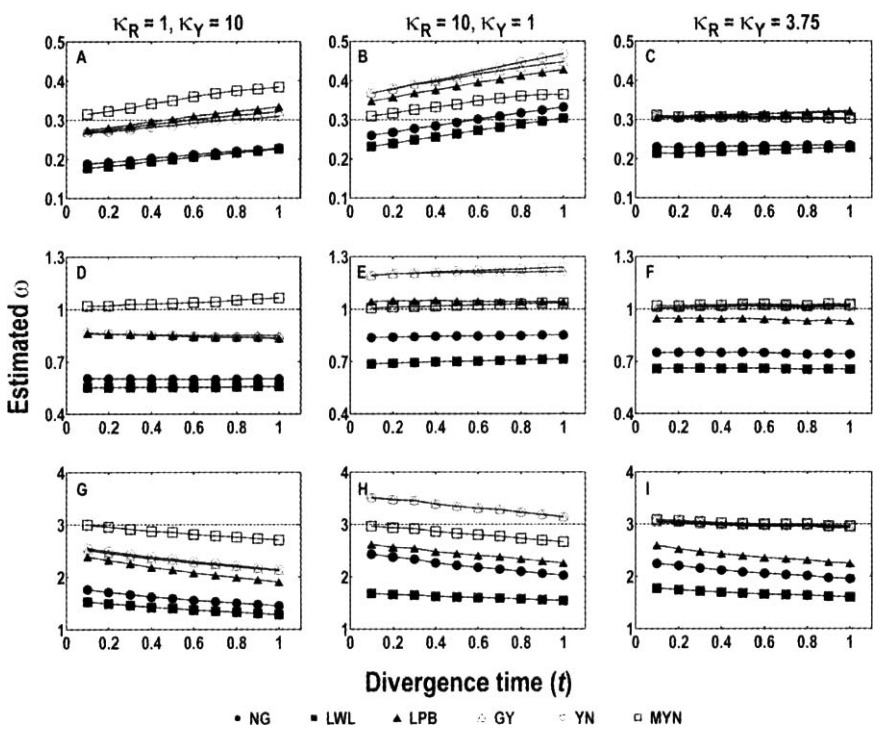

Fig. 4 Estimates of ω with six methods considering divergence time (t) varying from 0.1 to 1. Sequences are simulated with the human codon frequencies derived from human protein-coding genes. Three different combinations of κ_R and κ_Y are examined: $\kappa_R = 1$, $\kappa_Y = 10$ (A, D, G); $\kappa_R = 10$, $\kappa_Y = 1$ (B, E, H); $\kappa_R = \kappa_Y = 3.75$ (C, F, I). $\omega = 0.3$ (A–C), $\omega = 1$ (D–F), and $\omega = 3$ (G–I) are considered as typical values for purifying selection, neutral mutation, and positive selection, respectively.

$\kappa_R < \kappa_Y$, overestimate ω when $\kappa_R > \kappa_Y$, and yield closer estimates when $\kappa_R = \kappa_Y$. As to MYN, the bias of estimated ω tends to become more serious as t increases to extremes when $\kappa_R \neq \kappa_Y$, whereas t has minor influence and closer estimates could be observed when $\kappa_R = \kappa_Y$.

Evaluation results

NG and LWL

NG considers all possible evolutionary pathways among compared DNA sequences and assumes that each nucleotide is substituted by any other at equal rate (κ_R=κ_Y=1) when counting sites and substitutions. It adopts the Jukes-Cantor's one-parameter formula (*17*) to correct for multiple substitutions. Since transitions are more likely to occur than transversions, NG often underestimates transition/transversion rate ratio (κ) and thus the number of synonymous sites (S), which results in overestimation of Ks and underestimation of ω. This phenomenon can be observed in our simulation results, which was also found by Yang and Nielsen (*20*).

LWL classifies sites and substitutions as i-fold degenerate sites (i = 0, 2, 4) (three-fold degenerate sites, ATT, ATC, and ATA, are considered as two-fold ones). It considers unequal rates between transitional and transversional changes only when counting substitutions, but equal rates when counting sites. In detail, LWL assumes that two-fold degenerate sites are one-third synonymous and two-thirds nonsynonymous with Equations 1 and 2:

$$\mathrm{Ka} = \frac{L_0K_0 + L_2B_2}{L_0 + 2L_2/3} \quad (1)$$

$$\mathrm{Ks} = \frac{L_4K_4 + L_2A_2}{L_2/3 + L_4} \quad (2)$$

where L_i is the number of i-fold degenerate sites, and A_i and B_i are the numbers of transitional and transversional substitutions per i-fold degenerate site (i = 0, 2, 4), respectively. Hence, the total number (K_i) of substitutions per i-fold degenerate site is formulated as $K_i = A_i + B_i$. Although the Kimura's two-parameter formulas (*15*) are used for correction of multiple substitutions, LWL performs similarly to NG, since the number of substitutions in most cases is less than that of sites and thus the influence of κ on substitutions is not stronger than that on sites (LWL considers the transition/transversion bias only when counting substitutions).

Interestingly, it seems that NG and LWL, with increasing t, tend to give better estimates for purifying selection but biased estimates for positive selection, whereas t has no influence on them for neutral mutation (Figure 4). To explain this result, we derived an approximate formula for ω = Ka/Ks $\approx$ $(N_d/N)/(S_d/S) = (S/N)\times(N_d/S_d)$ (the symbol of "$\approx$" is due to the absence of correcting for multiple substitutions). Therefore, ω is composed of two parts: S/N, which is always underestimated, arising from the assumption of $\kappa_R = \kappa_Y = 1$ when counting sites by NG and LWL; N_d/S_d, which is related to t since an increase in t leads to more substitutions. For purifying selection, synonymous substitutions are more likely to occur than nonsynonymous ones. Therefore, small t tends to give rise to synonymous substitutions with only one difference, whereas an increase in t may result in more differences between two compared codons and thus provoke not only synonymous substitutions but also nonsynonymous ones according to different evolutionary pathways. Hence, for purifying selection, the value of N_d/S_d is on the rise as t increases, which can cancel the underestimation of S/N and thus lead to better estimates of ω for large t than for small t. In a similar way, the value of N_d/S_d for positive selection is on the decrease as t increases, which leads to the underestimation of ω; for neutral mutation, the value of N_d is close to that of S_d since synonymous and nonsynonymous substitutions per site occur with equal frequency and therefore ω seems to be nearly constant. These theoretical results are consistent well with the data found in Figure 4.

LPB

LPB, proposed as a modification of LWL, corrects for the bias in counting sites by using different formulas for Ka and Ks estimations (which is the only difference between LWL and LPB) with Equations 3 and 4:

$$\mathrm{Ka} = A_0 + \frac{L_0B_0 + L_2B_2}{L_0 + L_2} = \frac{L_0K_0 + L_2(A_0 + B_2)}{L_0 + L_2} = \frac{L_0(A_0 + B_0) + L_2(A_0 + B_2)}{L_0 + L_2} \quad (3)$$

$$\mathrm{Ks} = B_4 + \frac{L_2A_2 + L_4A_4}{L_2 + L_4} = \frac{L_4K_4 + L_2(A_2 + B_4)}{L_2 + L_4} = \frac{L_4(A_4 + B_4) + L_2(A_2 + B_4)}{L_2 + L_4} \quad (4)$$

LPB considers that Ka comprises two parts: the transitional nonsynonymous substitution rate A_0 and the transversional nonsynonymous substitution rate $(L_0B_0+L_2B_2)/(L_0+L_2)$. Likewise, Ks comprises the transitional synonymous substitution rate B_4 and the transversional synonymous substitution rate $(L_2A_2+L_4A_4)/(L_2+L_4)$. Based on these modifications, LPB improves the performance of LWL for most of the parameter combinations observed in simulations.

It can be observed from our comparative results that LPB tends to overestimate ω for purifying selection and to underestimate ω for positive selection. This result can be explained by assuming Ka = K_0 and Ks = K_4 at the perfect condition, since substitutions at nondegenerate sites are all nonsynonymous and those at four-fold degenerate sites are all synonymous (*10*). We reformulated the equations of Ka and Ks and found that the weighted average of K_0 and A_0+B_2 over 0- and 2-fold degenerate sites is considered as Ka (Equation 3) and the weighted average of K_4 and A_2+B_4 over 2- and 4-fold degenerate sites is Ks (Equation 4).

Let us first examine purifying selection. Transversions at two-fold degenerate sites can lead to synonymous substitutions (for example, CGG to AGG), whereas those at nondegenerate sites cannot. Since synonymous substitutions are more likely to occur than nonsynonymous ones for purifying selection, the value of B_0 is less than that of B_2, which leads to $K_0 = (A_0+B_0) < (A_0+B_2)$. In addition, the value of A_4 is greater than that of A_2 due to the fact that synonymous substitutions occur with higher possibilities at four-fold degenerate sites than two-fold ones. As a result, $K_4 = (A_4+B_4) > (A_2+B_4)$. Therefore, we can conclude that LPB overestimates ω for purifying selection, arising from Ka > K_0 and Ks < K_4. For positive selection, LPB underestimates Ka and overestimates Ks in a similar way, resulting in the underestimation of ω. We can see that this theoretical conclusion agrees well with simulation results for most of the parameter combinations examined.

GY, YN, and MYN

GY, based on a codon-based model, takes account of more features of DNA sequence evolution (such as transition/transversion rate bias and nucleotide/codon frequency bias) and calculates Ka and Ks by maximum likelihood estimation. YN, a simplified version of GY, adopts the Hasegawa-Kishino-Yano model (*16*) that also considers these evolutionary features and thus gives a close approximation of the maximum-likelihood method. To allow for more features of sequence evolution, MYN exploits the Tamura-Nei model (*18*) and uses two different ratios of transitional rate between purines (κ_R) and between pyrimidines (κ_Y) over the transversional rate when counting sites and substitutions. As a whole, these three methods perform better than NG, LWL, and LPB, while MYN improves the performance of YN for most of the parameter combinations (*14*).

However, we cannot conclude that which one of them is more accurate than other methods since our simulations are merely approximate and all methods may more or less give biased results for at least some parameter settings. We summarized the above analyses for these six methods in Table 2. In addition, there still have other methods (*21–24*) that are not included in our study. For example, a method similar to GY was proposed by Muse and Gaut (*24*), and some modified versions of LWL or LPB were improved by subdividing two-fold degenerate sites and substitutions, taking account of transition/transversion rate bias in counting sites, correcting for Arginine (ATT, ATC, and ATA), and so on.

Table 2 Mutation Models and Evolutionary Features in Different Methods

Method	Mutation model	Transition/transversion		Codon/nucleotide frequency
		site[#1]	substitution[#2]	
NG	Jukes-Cantor	$\kappa_R = \kappa_Y = 1$	$\kappa_R = \kappa_Y = 1$	equal
LWL	Kimura	$\kappa_R = \kappa_Y = 1$	$\kappa_R = \kappa_Y$	equal
LPB	Kimura	–*	–*	equal
GY	Codon-based	$\kappa_R = \kappa_Y$	$\kappa_R = \kappa_Y$	unequal
YN	Hasegawa-Kishino-Yano	$\kappa_R = \kappa_Y$	$\kappa_R = \kappa_Y$	unequal
MYN	Tamura-Nei	$\kappa_R \neq \kappa_Y$	$\kappa_R \neq \kappa_Y$	unequal

[#] κ_R and κ_Y are assumed by different methods: [#1]in the step of counting sites and [#2]in the step of counting substitutions. *LPB has no specific definition of synonymous and nonsynonymous sites or substitutions.

Discussion

It can be found from our results that incorporating more features of sequence evolution (such as transition/transversion bias and nucleotide/codon frequency bias) into Ka and Ks estimations could accurately capture more reliable estimates among protein-coding sequences. Although it is still hard to accommodate the trade-off between considering more parameters (evolutionary features) and avoiding over-parameterization (*25*), and simple methods (such as NG) are more suitable for short sequences, methods taking more evolutionary features into account should be the first choice to yield estimates with high quality and accuracy.

However, it should also be noted that all methods may one way or another give rise to biased results for at least some parameter combinations. How can we obtain the most reliable estimates of Ka and Ks? As mentioned above, these methods adopt different nucleotide substitution or mutation models, leading to diverse estimates of evolutionary distance (*19*). In addition, since the amount and the degree of sequence substitutions vary among datasets, a single model or a single method is not adequate for Ka and Ks calculations. As a consequence, model selection, that is, choosing a best-fit model according to compared sequences when estimating Ka and Ks, becomes critical for capturing appropriate evolutionary information (*26*). Therefore, implementation of different mutation models in a framework of maximum likelihood could help us include as many features as needed in Ka and Ks estimations, which accordingly needs the Akaike Information Criterion or the Bayesian Information Criterion (*27*) as a measure of fitness between models and data.

Materials and Methods

Simulated sequences were generated from hypothetical common ancestral sequences. Each codon of the common ancestral sequences was randomly chosen from 64 codons (except stop codons) according to codon frequencies. In this study, we considered three sets of codon frequencies derived from three empirical datasets: (1) equal codon frequencies, that is, each sense codon frequency is 1/(64−the number of stop codons) (*13*); (2) human codon frequencies deduced from 39,420 human protein-coding genes from the ENSEMBL database (Release 35; ref. *28*); and (3) rice codon frequencies retrieved from 19,079 rice protein-coding genes (*29*).

In addition to codon frequencies, other parameters were set in simulations, including sequence length, divergence time (t), two ratios of transitional rate between purines (κ_R) and between pyrimidines (κ_Y) to transversional rate, and the selective strength (ω = Ka/Ks). Although ω varies from gene to gene, ω = 0.3, 1, and 3 can be regarded as "typical values" for negative selection, neutral mutation, and positive selection, respectively (*3, 13, 30*), which could be observed from real datasets. To accurately examine the effect of one parameter and avoid stochastic errors arising from other factors, we simulated sequences with 2,000,000 codons.

To compare the accuracies of Ka and Ks estimations with different methods, we estimated expected Ks and Ka values by counting the numbers of synonymous and nonsynonymous sites of ancestral sequences and then using the formulas Ks = 3×(S+N)×t/ (S+ω×N) and Ka = ω×Ks. Considering that simulated sequences have different expected values of Ka and Ks, we used the formula 100% × [(estimated value) − (expected value)]/(expected value) to calculate percentage errors for a better display of relative biases between estimated and expected values.

For Ka and Ks estimations, we used six different methods: NG, LWL, LPB, GY, YN, and MYN, which have been implemented in our software KaKs_Calculator (prepared for submission). For error-checking, we also compared the results estimated by KaKs_Calculator with those by other tools, such as PAML (*31*), in which NG, GY, and YN were implemented.

Acknowledgements

This work was supported by the Ministry of Science and Technology of China (Grant No. 2001AA231061) and the National Natural Science Foundation of China (Grant No. 30270748) awarded to JY. We thank Mr. Jun Li for valuable discussion.

Authors' contributions

ZZ performed computer simulations to generate sequences, carried out the comparative analysis, and drafted the manuscript. JY supervised the research and revised the manuscript. Both authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Kimura, M. 1983. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK.
2. Gillespie, J.H. 1991. *The Causes of Molecular Evolution.* Oxford University Press, Oxford, UK.
3. Li, W.H. 1997. *Molecular Evolution.* Sinauer Associates, Sunderland, USA.
4. Nekrutenko, A., *et al.* 2002. The Ka/Ks ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* 12: 198-202.
5. Pal, C., *et al.* 2006. An integrated view of protein evolution. *Nat. Rev. Genet.* 7: 337-348.
6. Hurst, L.D. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. *Trends Genet.* 18: 486.
7. Yang, Z. and Bielawski, J.P. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15: 496-503.
8. Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3: 418-426.
9. Li, W.H., *et al.* 1985. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2: 150-174.
10. Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 36: 96-99.
11. Pamilo, P. and Bianchi, N.O. 1993. Evolution of the *Zfx* and *Zfy* genes: rates and interdependence between the genes. *Mol. Biol. Evol.* 10: 271-281.
12. Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11: 725-736.
13. Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17: 32-43.
14. Zhang, Z., *et al.* 2006. Computing Ka and Ks with a consideration of unequal transitional substitutions. *BMC Evol. Biol.* 6: 44.
15. Kimura, M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *J. Mol. Evol.* 16: 111-120.
16. Hasegawa, M., *et al.* 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22: 160-174.
17. Jukes, T.H. and Cantor, C.R. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. Munro, H.N.), pp. 21-123. Academic Press, New York, USA.
18. Tamura, K. and 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.
19. Muse, S.V. 1996. Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* 13: 105-114.
20. Yang, Z. and Nielsen, R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46: 409-418.
21. Comeron, J.M. 1995. A method for estimating the numbers of synonymous and nonsynonymous substitutions per site. *J. Mol. Evol.* 41: 1152-1159.
22. Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. *J. Mol. Evol.* 40: 190-226.
23. Tzeng, Y.H., *et al.* 2004. Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 21: 2290-2298.
24. Muse, S.V. and Gaut, B.S. 1994. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. *Mol. Biol. Evol.* 11: 715-724.
25. Lio, P. and Goldman, N. 1998. Models of molecular evolution and phylogeny. *Genome Res.* 8: 1233-1244.
26. Sullivan, J. and Joyce, P. 2005. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. S.* 36: 445-466.
27. Posada, D. and Buckley, T.R. 2004. Model selection and model averaging in phylogenetics: advantages of akaike information criterion and bayesian approaches over likelihood ratio tests. *Syst. Biol.* 53: 793-808.
28. Hubbard, T., *et al.* 2005. Ensembl 2005. *Nucleic Acids Res.* 33: D447-453.
29. Yu, J., *et al.* 2005. The genomes of *Oryza sativa*: a history of duplications. *PLoS Biol.* 3: e38.
30. Messier, W. and Stewart, C.B. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385: 151-154.
31. Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13: 555-556.