How Do Variable Substitution Rates Influence Ka and Ks Calculations?

Dapeng Wang[1,2], Song Zhang[1,2,3], Fuhong He[1,2], Jiang Zhu[1,2], Songnian Hu[1], and Jun Yu[1,3]*

[1] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China; [2] Graduate University of Chinese Academy of Sciences, Beijing 100049, China; [3] Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100080, China.

*Corresponding author. E-mail: junyu@big.ac.cn

DOI: 10.1016/S1672-0229(08)60040-6

The ratio of nonsynonymous substitution rate (Ka) to synonymous substitution rate (Ks) is widely used as an indicator of selective pressure at sequence level among different species, and diverse mutation models have been incorporated into several computing methods. We have previously developed a new γ-MYN method by capturing a key dynamic evolution trait of DNA nucleotide sequences, in consideration of varying mutation rates across sites. We now report a further improvement of NG, LWL, MLWL, LPB, MLPB, and YN methods based on an introduction of gamma distribution to illustrate the variation of raw mutation rate over sites. The novelty comes in two ways: (1) we incorporate an optimal gamma distribution shape parameter *a* into γ-NG, γ-LWL, γ-MLWL, γ-LPB, γ-MLPB, and γ-YN methods; (2) we investigate how variable substitution rates affect the methods that adopt different models as well as the interplay among four evolutional features with respect to Ka/Ks computations. Our results suggest that variable substitution rates over sites under negative selection exhibit an opposite effect on ω estimates compared with those under positive selection. We believe that the sensitivity of our new methods has been improved than that of their original methods under diverse conditions and it is advantageous to introduce novel parameters for Ka/Ks computation.

Key words: substitution rate, approximate method, gamma distribution, Ka, Ks

Introduction

One of the important parameters for molecular evolutionary analyses is the estimation of the synonymous (Ks) and nonsynonymous (Ka) nucleotide substitution rates, which are respectively defined as the number of synonymous substitutions per synonymous site and the number of nonsynonymous substitutions per nonsynonymous site per year or per generation. It is commonly accepted that Ka>Ks, Ka=Ks, and Ka<Ks generally indicate positive selection, neutral mutation, and negative selection, respectively (*1*, *2*). There are multifarious methods for estimating Ka and Ks on the basis of various substitution models, which are categorized into two essential types: approximate methods and maximum likelihood ones. In practice, these methods should be applied cautiously and simple conclusions are not easily drawn when only one method is adopted (*3*). Therefore, it is necessary for us to continue developing diversified models to accurately calculate Ka and Ks.

Since both approximate and maximum likelihood methods usually yield similar estimates based on the same hypothesis (*2*, *4*) and the latter are often time-consuming (*5*), we only focus on the approximate methods for our analyses. Most existing methods, such as NG (Nei-Gojobori) (*6*), LWL (Li-Wu-Luo) (*7*), MLWL (a modified LWL method) (*8*), LPB (Li-Pamilo-Bianchi) (*9*, *10*), MLPB (a modified LPB method) (*8*), YN (Yang-Neilsen) (*5*), and MYN (a modified YN method) (*11*), consider three significant dynamic features of evolving DNA sequences: transition/transversion rate bias, nucleotide frequency bias, and unequal transitional substitution, but omit another substantial character—unequal substitution rates across sites. In fact, rate variation among nucleotide sites is commonly observed, due to the functional restraint of amino acids at the active centers of proteins (*1*, *2*). In particular, this is true for protein-coding genes where the three codon positions have different functional constraints for nucleotide substitutions (*1*, *12*). Since γ-distribution has been widely used to illustrate the characteristics of nucleotide mutation rate (*13–16*), especially in the field of estimating sequence divergence (*6*, *14*, *17–21*), we have developed a γ-MYN method (*22*) by introducing γ-distribution into MYN method (*11*), and observed that the performance of the new method is better than that of the original one under certain conditions. In this paper, we bring this assumption into other existing methods so that the series of new γ-methods are denoted as γ-NG, γ-LWL, γ-MLWL, γ-LPB, γ-MLPB, and γ-YN. We focus on the performance evaluation of these new methods in combination with properties of various parameters and dynamic features of evolving DNA sequences as well as their influences on Ka and Ks calculations. The descriptions of symbols used in this paper are shown in **Table 1**.

Results and Discussion

Effect of γ-distribution on various methods

On the assumption that the rate of nucleotide substitution approximately follows the gamma distribution, we have supplemented seven methods: γ-NG, γ-LWL, γ-MLWL, γ-LPB, γ-MLPB, γ-YN, and γ-MYN (*22*). Since γ-MLPB performs the same as γ-LPB does (**Tables 2** and **S1**; data not shown), we chose γ-LPB for our analyses. We plotted the percentage errors for Ka and Ks, and estimated ω against κ_{R} for different expected values, using rice codon frequencies in three conditions of expected ω=0.3, 1, and 3, respectively (**Figures 1–3** and **S1–S6**).

Let us examine the characteristics of these plots in general. Among them, the curves yielded from γ-NG and γ-LWL remain nearly horizontal regardless the variables Ka, Ks, or ω (Figures 1–3 and S1–S6). When we examined Ka and Ks, the trends from γ-MLWL, γ-LPB, and γ-YN showed the opposite directions, increasing for Ka and decreasing for Ks (Figures S1–S6). The trend from γ-MYN seems distinct from all the other methods (Figures 1–3 and S1–S6). From above observations, we categorized these six methods into three categories: (1) γ-NG and γ-LWL; (2) γ-MLWL, γ-LPB, and γ-YN; and (3) γ-MYN, according to their similar tendencies as key parameter varies. We believe that the reason for such tendencies is related to their underlying models; as we know, γ-MLWL, γ-LPB, and γ-YN consider transition/transversion rate bias, γ-MYN takes unequal transitional substitution (between the two purines, or the two pyrimidines), while both γ-NG and γ-LWL leave out the major dynamic features of evolving DNA sequences utilized by other methods.

We now investigate how the diversified values of shape parameter a affect the performances of various methods. Mathematically, when $a \to \infty$, γ-series methods are reduced to their corresponding conventional methods. For example, as $a \to \infty$, γ-LWL→LWL. Naturally, we denoted $a \to \infty$ as $a = \infty$ for simplicity and chose six values (0.2, 0.6, 1, 4, 20, and ∞) as typical a values. Here we did not show the results related to conditions of a=0.2 and $a=\infty$ for two reasons. First, the curves of a=0.2 always extend out of the normal range in comparison with the expected outlines (data not shown) as these cases may not be meaningful for arithmetic applications. Second, the curves of a=20 and $a=\infty$ perform so similar that we are unable to distinguish them (data not shown), therefore we used one of them, a=20, not a=0.2 and $a=\infty$. In Figures 1–3 and S1–S6, we

Table 1 Symbols used in this paper

Symbol	Description
S	Number of synonymous sites
N	Number of nonsynonymous sites
Ks	Synonymous substitution rate
Ka	Nonsynonymous substitution rate
ω	Estimator of selective pressure, ω=Ka/Ks
$\mathrm{S_d}$	Number of synonymous substitutions
$\mathrm{N_d}$	Number of nonsynonymous substitutions
t	Divergence time between two sequences
a	The shape parameter of gamma distribution
α	Transitional rate
α_1	Transitional rate between purines
α_2	Transitional rate between pyrimidines
β	Transversional rate
κ	Ratio of transitional rate/transversional rate
κ_{R}	Ratio of transitional rate between purines to transversional rate, $\kappa_{\mathrm{R}}=\alpha_1/\beta$
κ_{Y}	Ratio of transitional rate between pyrimidines to transversional rate, $\kappa_{\mathrm{Y}}=\alpha_2/\beta$
g_{N}	Frequency of nucleotide N, N∈[T, C, A, G]
g_{R}	$g_{\mathrm{R}} = g_{\mathrm{A}} + g_{\mathrm{G}}$
g_{Y}	$g_{\mathrm{Y}} = g_{\mathrm{T}} + g_{\mathrm{C}}$

Table 2 The optimal values of gamma distribution shape parameter a based on a combination of nine terms and seven methods

Condition	Term	a values						
		γ-NG	γ-LWL	γ-MLWL	γ-LPB	γ-MLPB	γ-YN	γ-MYN
ω=0.3	Ka	0.6	0.6	∞	∞	∞	∞	20
	Ks	∞	∞	4	4	4	20	∞
	ω	∞	∞	4	1	1	4	20
ω=1	Ka	1	1	4	20	20	20	4
	Ks	∞	∞	∞	4	4	4	20
	ω	∞	∞	∞	∞	∞	∞	∞
ω=3	Ka	4	4	4	4	4	4	4
	Ks	∞	∞	∞	∞	∞	4	4
	ω	0.6	0.2	0.6	1	1	∞	∞

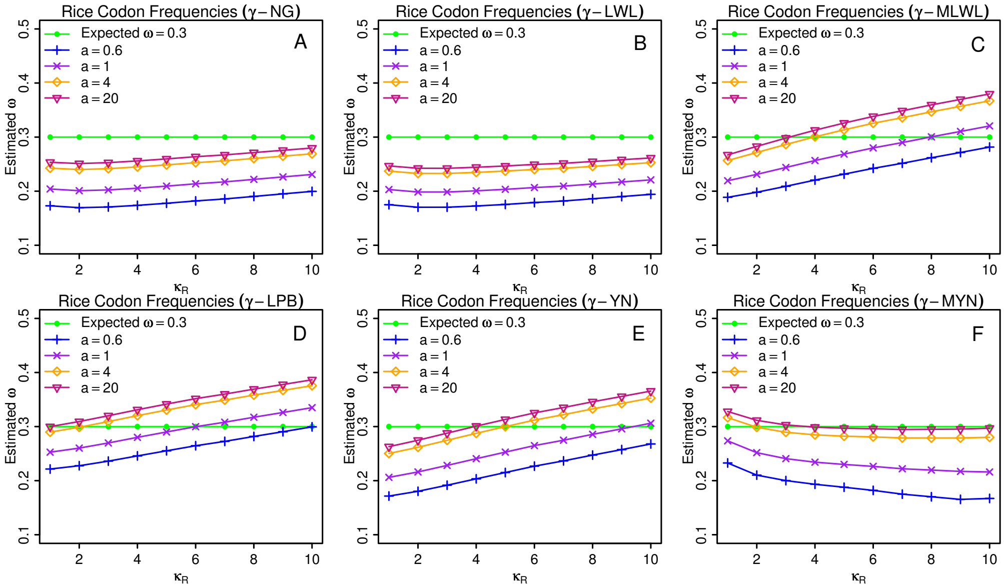

Figure 1 Average ω estimates under the condition of expected ω=0.3. We plotted average ω estimates over 2,000 pairs of sequences based on γ-NG, γ-LWL, γ-MLWL, γ-LPB, γ-YN, and γ-MYN, when κ_Y=3.75 and κ_R varies from 1 to 10, under the condition of expected ω=0.3.

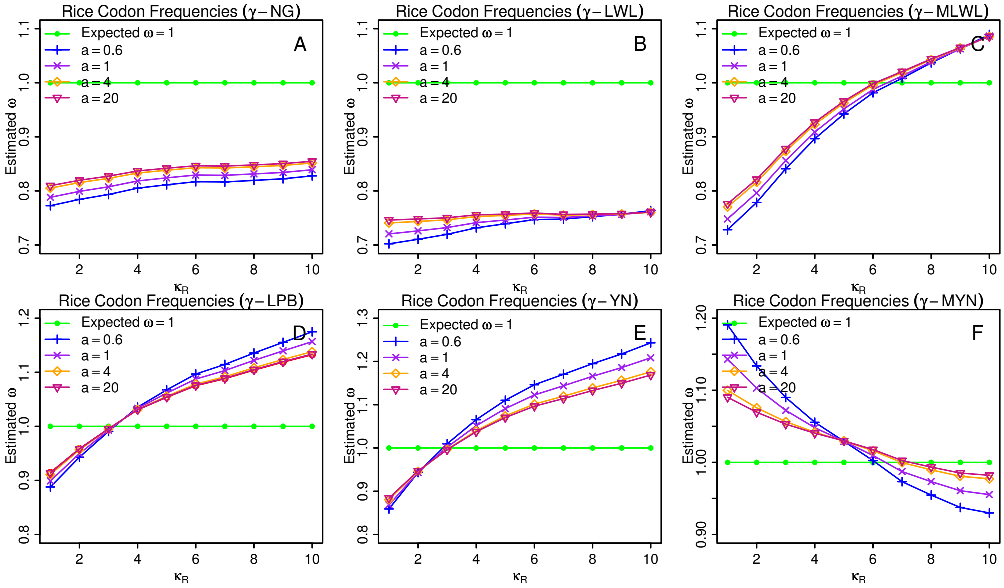

Figure 2 Average ω estimates under the condition of expected ω=1.

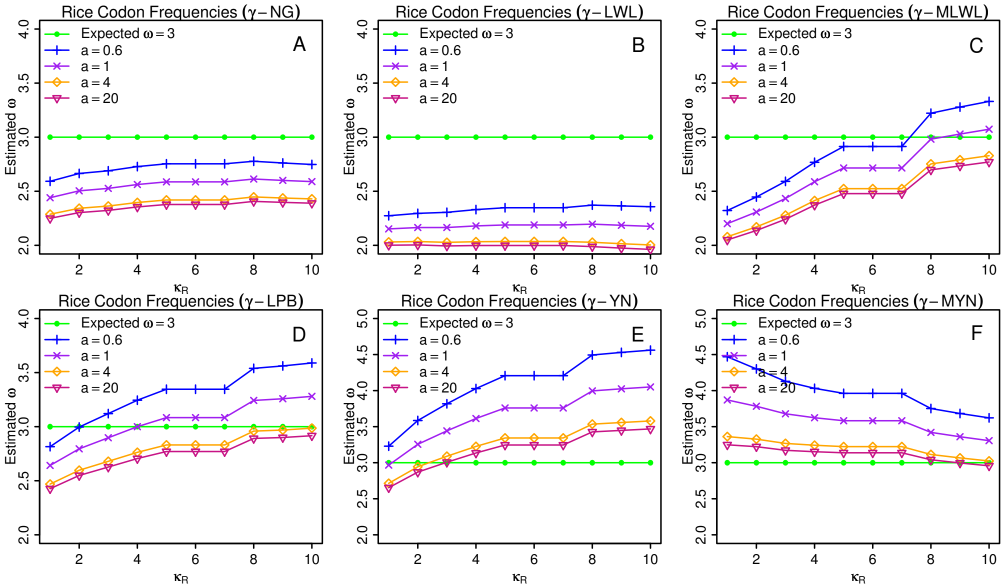

Figure 3 Average ω estimates under the condition of expected ω=3.

noticed that most of the curves remain parallel as a varies with minor exceptions in Figure 2. We have a few interesting observations. First, each curve rises in parallel as a decreases when Ka and Ks are examined, regardless whether ω=0.3, 1, or 3. Even though our findings on expected ω=3 are consistent with above observations when ω is examined, the results when expected ω=0.3 are opposite under most of the other conditions. We believe that it is attributable to the distributions of the curves under two other assumptions: expected ω=0.3 and ω=3 (Figures 1 and 3). Interestingly, when expected ω=1, each curve seems to rotate around the center in each panel (Figure 2) when ω is examined. Next, when expected ω=0.3, ω changes lie on those of Ks, due to the fact that Ks is more sensitive to the changes of a than Ka. When expected ω=3, ω changes depend on Ka as Ka is more sensitive to a changes. When expected ω=1, a changes have less impact on ω, due to the fact that Ka and Ks have similar sensitivity to a changes. Combining above observations, we conclude that larger values of Ka and Ks are more sensitive to the changes of a.

The optimal values of gamma distribution shape parameter *a*

We computed the optimal indexes (see Materials and Methods) for optimal values of a under various conditions (Table S1) and found the minimal values in each column, whose corresponding a values are considered as optimal (Table 2). To study the implication, we divided a into three categories (*1*, *2*) according to the shapes of γ-distribution (**Figure 4**): (1) when a<1, the distribution indicates that most of the sites have very low substitution rates despite the existence of a few sites with higher substitution rates; (2) when a>1, the distribution shows that the majority of the sites have intermediate rates around 1, except the fact that some sites may exhibit extreme rates (very low or high); (3) when a goes to the infinity, the distribution becomes a simpler type that all sites have the same rate. Now we only discuss the term ω in combination with Table 2. When the positive and negative selection forces balance each other (neutral mutation), all sites evolve in the same rate regardless what methods

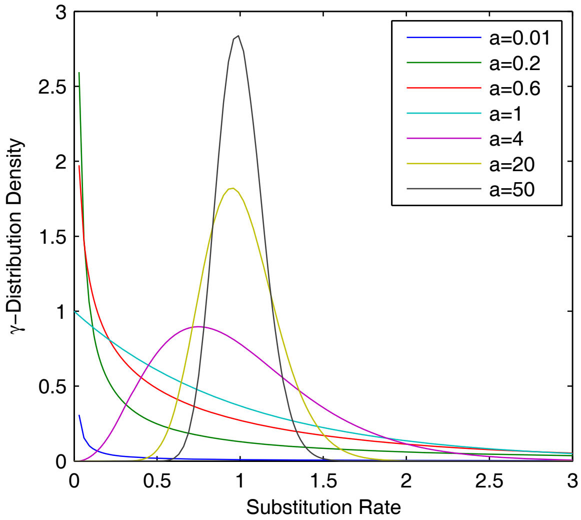

Figure 4 γ-distribution densities as a function of substitution rates at various a values of 0.01, 0.2, 0.6, 1, 4, 20, and 50.

were actually used. When γ-NG, γ-LWL, and γ-MLWL are examined, a value decreases with the increasing selective pressure varying from 0.3 to 3. This indicates that significant increase in selective pressure makes more sites evolve in very low rate. However, we found slightly opposite effects in γ-YN and γ-MYN, perhaps due to their shared consideration in nucleotide frequency bias (codon frequency bias) and the complex interplay between nucleotide frequency and variable substitution rates across sites. Another interesting observation is that the pattern of rate variation at sites holds the line under the conditions of both ω=0.3 and ω=3, when γ-LPB is examined.

Effect of codon frequencies

To examine the influence of codon frequencies on the capability of our new methods, we simulated hypothetical common ancestral sequences on the basis of three datasets: equal, human, and rice codon frequencies. We estimated the performance of our new methods at their optimal values of a under three conditions of ω=0.3, ω=1, and ω=3, using three sets of codon frequencies (**Figure 5A–I**). As a whole, different codon frequencies have little influence on the performance of our new methods. We also found that their performances under human codon frequencies are similar to those under rice codon frequencies but not under equal codon frequencies.

Effect of *t*

To examine the effect of divergence time based on our new methods, we plotted estimated ω against t (from 0.1 to 1), using rice codon frequencies (**Figure 6**). To measure the robustness of the methods, we focused on

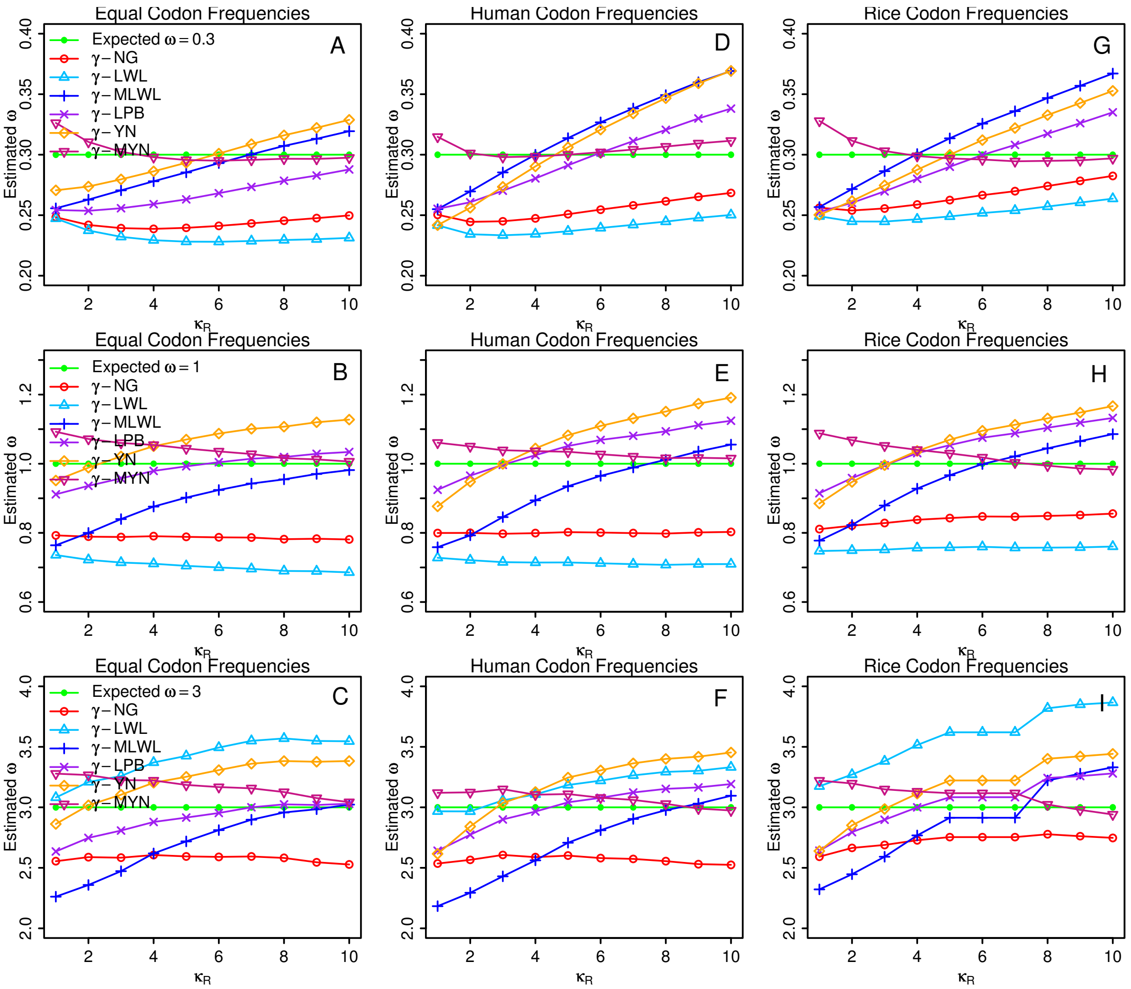

Figure 5 Average ω estimates based on the six methods under three different codon frequencies, when κ_{Y}=3.75 and κ_{R} varies from 1 to 10. The codon frequencies used are: equal (A, B, C), human (D, E, F), and rice (G, H, I). ω=0.3 (A, D, G), ω=1 (B, E, H), and ω=3 (C, F, I) stand for purifying selection, neutral mutation, and positive selection, respectively. The values of a used in the six methods are listed in Table 2.

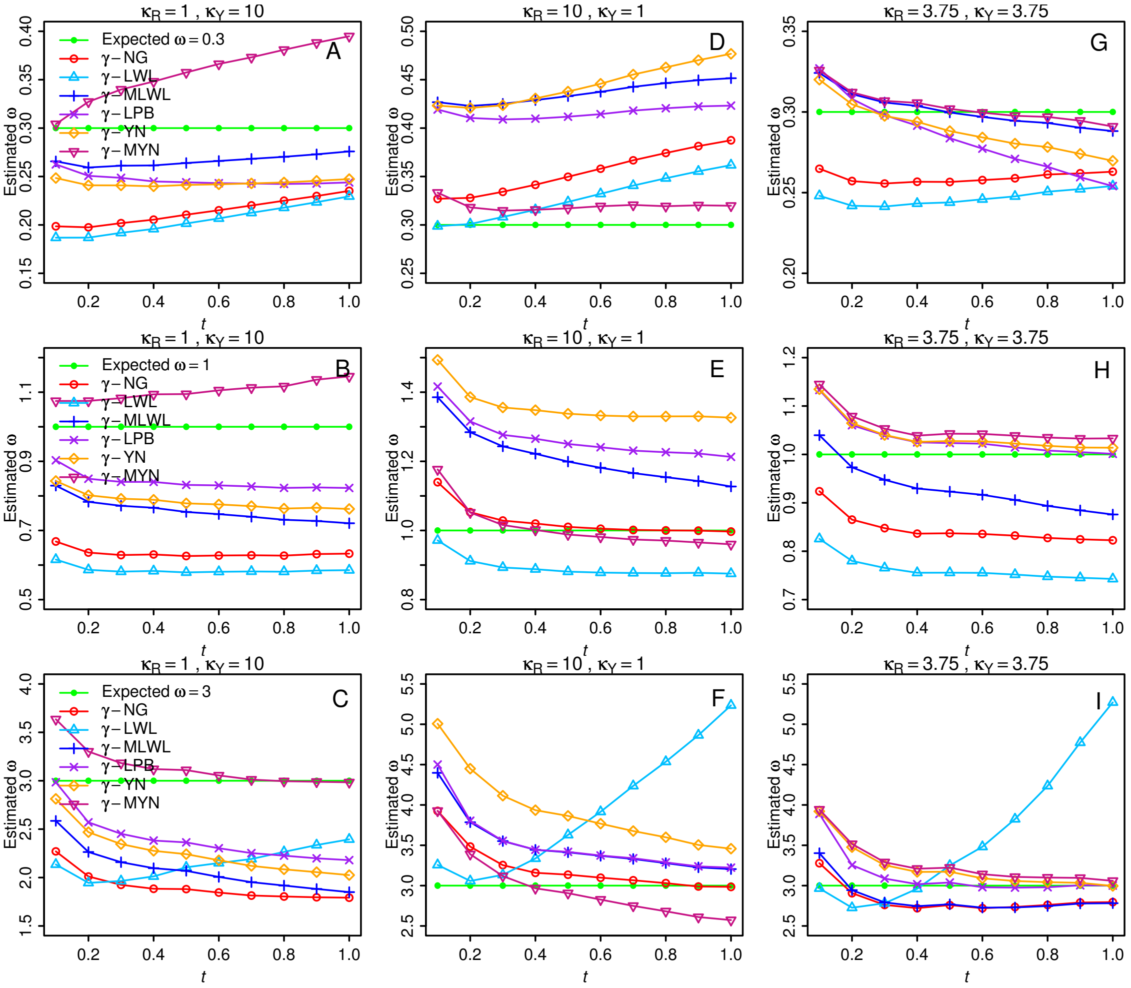

Figure 6 Average ω estimates based on the six methods with the consideration of divergence time (t) that varies from 0.1 to 1. We considered the typical values for purifying selection, neutral mutation, and positive selection as ω=0.3 (A, D, G), ω=1 (B, E, H), and ω=3 (C, F, I), respectively. Three different combinations of $\kappa_{\rm R}$ and $\kappa_{\rm Y}$ were examined: $\kappa_{\rm R}$=1, $\kappa_{\rm Y}$=10 (A, B, C); $\kappa_{\rm R}$=10, $\kappa_{\rm Y}$=1 (D, E, F); $\kappa_{\rm R}$=$\kappa_{\rm Y}$=3.75 (G, H, I). The values of a used for the six methods are listed in Table 2.

three extreme cases: (1) $\kappa_{\rm R}$=1, $\kappa_{\rm Y}$=10; (2) $\kappa_{\rm R}$=10, $\kappa_{\rm Y}$=1; and (3) $\kappa_{\rm R}$=$\kappa_{\rm Y}$=3.75. In general, most of them do not change much as t increases; it is a sign for robustness. One exception is γ-LWL when the expected ω is 3 and when $\kappa_{\rm R}$=10, $\kappa_{\rm Y}$=1, and $\kappa_{\rm R}$=$\kappa_{\rm Y}$=3.75. The fact suggests that γ-LWL is less robust when t approaches the extreme. We thought that the divergence time t is the major factor. However, γ-LWL performs well when $\kappa_{\rm R}$=1, $\kappa_{\rm Y}$=10, and the expected ω=3.

Effects of other parameters

We are aware of other parameters used for arithmetical estimation (*5*, *11*) but paid less attention to them. For S% (the percentage of synonymous sites in a sequence), we found that γ-NG, γ-LWL, and γ-MLWL do not change the estimation of S% much but γ-LPB, γ-YN, and γ-MYN always overestimate S% to different extent (data not shown). In terms of sequence length, an increase often induces biases (*11*). Since we chose an average sequence length of 400 codons for the analyses, we believe that our new methods should maintain their advantages when sequence length changes.

Testing real data

We utilized three mammalian homologous gene sets to verify the efficiency of these new methods. Plotting the distributions of $\kappa_{\rm R}-\kappa_{\rm Y}$ in three individual datasets and one pooled dataset (**Figure 7**), we found that the pooled dataset represents reasonably the three raw orthologous datasets and has sufficient gene

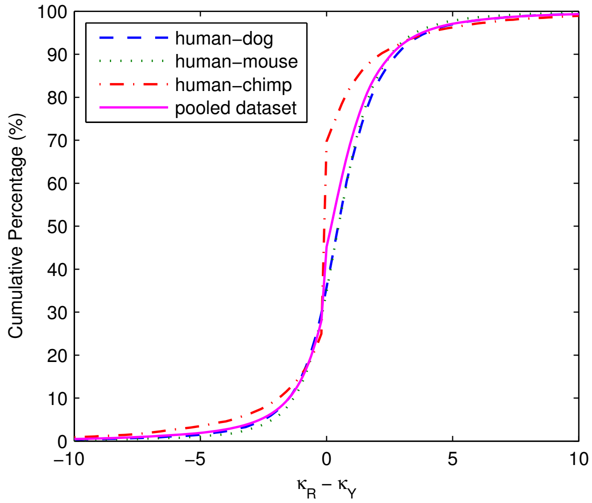

Figure 7 Cumulative percentage of $\kappa_R - \kappa_Y$ for human-dog, human-mouse, and human-chimp orthologs and a pooled dataset at a bin size of 0.2.

pairs falling in each interval of $\kappa_R - \kappa_Y$. Subsequently, we only dealt with the pooled data and analyzed S%, Ka, Ks, and ω in four intervals of $\kappa_R - \kappa_Y$ (**Table 3**). We also carefully selected three values (−0.5, 0.5, and 1.5) as segmentation boundaries to obtain the four subintervals, when κ_Y=3.75: (1) κ_R=1, 2 and 3; (2) κ_R=4; (3) κ_R=5; and (4) κ_R=6, 7, 8, 9 and 10. As the majority of genes are driven by negative selection, we set a values according to the optimal values when ω=0.3 (Table 2) for the convenience of comparing the results from the real data with those from computer simulations (Figure 5).

We have the following observations. First, the new γ-methods seem not overestimate ω, as compared to their original methods, in accordance with our simulation results and theoretical analyses. Second, we observed some variations of the new methods in ω estimates; for instance, γ-MYN produces consistent results with our simulations (Figure 5A, D, G). In the case of $\kappa_R - \kappa_Y < -0.5$, when κ_R=1, 2, and 3, γ-MYN overestimates ω compared with other γ-methods and the values are 0.2521, 0.2376, 0.2813, 0.2900, 0.2653, and 0.2944 for γ-NG, γ-LWL, γ-MLWL, γ-LPB, γ-YN, and γ-MYN, respectively. When confined $\kappa_R - \kappa_Y \geq 1.5$ (κ_R= 6, 7, 8, 9 and 10), γ-MLWL, γ-YN and γ-LPB overestimate ω evidently but γ-MYN, γ-NG and γ-LWL do not, as the values are 0.2127, 0.1918, 0.2088, 0.1684, 0.1817, and

Table 3 Estimates of S%, Ka, Ks, and ω based on an aggregate of three datasets and twelve methods

Method	$\kappa_R - \kappa_Y < -0.5$				$-0.5 \leq \kappa_R - \kappa_Y < 0.5$			
	S%	Ka	Ks	ω	S%	Ka	Ks	ω
NG/γ-NG	23.63%	0.0624	0.3705	0.2521	23.73%	0.0637	0.3128	0.2672
LWL/γ-LWL	22.36%	0.0625	0.3717	0.2376	22.46%	0.0620	0.3133	0.2508
MLWL	27.59%	0.0641	0.3114	0.2888	26.33%	0.0628	0.2738	0.2824
LPB	27.81%	0.0664	0.3033	0.3171	28.52%	0.0650	0.2602	0.3385
YN	25.45%	0.0635	0.4271	0.2731	24.38%	0.0634	0.3811	0.2617
MYN	26.86%	0.0646	0.3993	0.2960	24.29%	0.0636	0.3954	0.2593
γ-MLWL	27.59%	0.0660	0.3401	0.2813	26.33%	0.0649	0.3010	0.2755
γ-LPB	28.36%	0.0759	0.4393	0.2900	28.95%	0.0754	0.3796	0.3124
γ-YN	25.51%	0.0653	0.4935	0.2653	24.43%	0.0659	0.4447	0.2545
γ-MYN	26.88%	0.0651	0.4098	0.2944	24.30%	0.0641	0.4083	0.2577
Method	$0.5 \leq \kappa_R - \kappa_Y < 1.5$				$\kappa_R - \kappa_Y \geq 1.5$			
	S%	Ka	Ks	ω	S%	Ka	Ks	ω
NG/γ-NG	23.95%	0.0840	0.4701	0.2107	24.02%	0.0515	0.4027	0.1817
LWL/γ-LWL	22.69%	0.0842	0.4702	0.2040	22.74%	0.0521	0.4035	0.1729
MLWL	27.21%	0.0857	0.4050	0.2343	28.61%	0.0537	0.3315	0.2198
LPB	27.64%	0.0887	0.3884	0.2607	28.32%	0.0557	0.3282	0.2343
YN	25.22%	0.0835	0.5622	0.2047	26.14%	0.0515	0.4566	0.1991
MYN	24.38%	0.0825	0.6360	0.1874	24.33%	0.0502	0.5768	0.1703
γ-MLWL	27.21%	0.0884	0.4456	0.2241	28.61%	0.0548	0.3608	0.2127
γ-LPB	28.33%	0.1025	0.5770	0.2235	28.94%	0.0610	0.4570	0.2088
γ-YN	25.29%	0.0863	0.6573	0.1941	26.19%	0.0526	0.5291	0.1918
γ-MYN	24.40%	0.0830	0.6578	0.1849	24.34%	0.0504	0.6002	0.1684

0.1729 for γ-MLWL, γ-YN, γ-LPB, γ-MYN, γ-NG, and γ-LWL, respectively. However, when $-0.5 \leq \kappa_{\mathrm{R}} - \kappa_{\mathrm{Y}} < 1.5$, simulation results showed that the performance of each method becomes similar. Finally, Ka estimates among all γ-methods are very similar except γ-LPB. The major distinction in ω estimation with various γ-methods lies in Ks estimates—the changes of ω are mostly attributable to those of Ks as Ks is more sensitive than Ka to the changes of a under negative selection. In conclusion, our findings largely agree with the simulation studies.

How does the consideration of variable substitution rates improve Ka/Ks calculation?

Let us first examine how parameter a in the γ-series of methods improves the original methods. As we know, overlooking the fact of rate variation among sites often results in underestimation of both the sequence distance and the transition/transversion rate ratio κ (both κ_{R} and κ_{Y}) (*2*). The ratio κ plays a key role in two necessary processes of both (1) estimating S and N and (2) generating a transition probability matrix for computing $\mathrm{S_d}$ and $\mathrm{N_d}$, and therefore ω = Ka/Ks $\approx (\mathrm{N_d}/\mathrm{N})/(\mathrm{S_d}/\mathrm{S})$, where the "$\approx$" is a result of the absence of correcting for multiple hits.

We next discuss three special cases. The case of purifying selection has been discussed previously (*22*), and the underestimated κ is used in the original methods that lead to underestimation of $\mathrm{S_d}/\mathrm{S}$ and overestimation of ω in contrast to our γ-series methods. In the case of positive selection, we would like to only discuss Ka ($\mathrm{N_d}/\mathrm{N}$) since nonsynonymous substitutions are more likely to occur than synonymous ones. As κ is positively related to substitution number between two codons, underestimation of κ gives rise to underestimation of $\mathrm{N_d}$. Since it is more likely that transitions between two codons are synonymous, primarily at the third codon positions, the underestimation of κ often leads to the underestimation of S and the overestimation of N. Therefore, underestimations of $\mathrm{N_d}/\mathrm{N}$ or ω can be attributable to an underestimated κ. In the case of neutral mutation where synonymous substitutions occur in the same probability as nonsynonymous ones, a decrease in κ leads to dithering of the curves, and the power of parameter a is related to κ_{R} (or κ_{Y}), so we recommend to use the less complex conventional methods. Our analyses are consistent with the results from both simulation (Figures 1–3) and real data (Tables 2 and 3).

Usage, performance, and program availability

We evaluate the performance of the new methods using the parameters representing various selection pressures, especially negative selection, and often consider all conditions and integrate various parameter settings into the algorithm (Table 2) by identifying the scope of ω using a traditional method (ω>1 or ω<1) and computing final ω using γ-method with a combination of selected parameters. In our previous study (*22*), we showed that the GY method (a popular maximum likelihood method) consumes more time than approximate methods do. We therefore recommend our new methods to be used in the cases when large amounts of data are to be analyzed. C++ programs implementing γ-series methods such as γ-NG, γ-LWL, γ-MLWL, γ-LPB, and γ-YN are included in KaKs_Calculator version 2.0, which is a software package updated from KaKs_Calculator version 1.0 (*23*).

Prospective

As methods for calculating the two kinds of *distances*, nonsynonymous substitution rate as Ka and synonymous substitution rate as Ks between protein-coding sequences have been developed and widely used in the field of molecular evolution, and different models have been introduced into emerging new methods. However, it is still surprising that results from real data tend to produce similar results despite the fact that various methods are applied in parallel (*2*). Although it was shown that different correlations between selective pressure and Ks can be drawn from different methods (*24*), the major conclusions when detecting positive selection are not usually changed. Is it true that the Ka/Ks argument is too weak to have the ability in detecting positive selection? We believe that it is not, especially not due to the methodology for Ka/Ks calculations. By using these methods, we are able to obtain average selection pressures in a way where individual genes are used as an object. If one needs to determine whether any individual genes are subjected to positive selection, the LRT (likelihood ratio test)-like methods (*25*) should be used and they tend to be more qualitative. In conclusion, the two methods (LRT-like methods and Ka/Ks methods) should be applied to the study of different outcomes, and they are neither the same nor mutually exclusive. Therefore, our attempts in improving Ka/Ks methods are not only meaningful but also will increase the sensitivity to detect positive selection, especially when new strategies [*e.g.* sliding window (*26–31*)] are sought out for better resolutions.

Conclusion

We compared γ-methods with their conventional counterparts by carrying out computer simulations and examining real data. As neglecting the variation of substitution rates across sites may reflect on biased estimates of Ka and Ks in these examined methods, our new γ-methods have minimal deviations under various conditions. We show that incorporating variable substitution rates into the calculation of Ka and Ks and their ratio ω often exhibits merits over their conventional counterparts when applied appropriately.

Materials and Methods

Overview of general steps

Our γ-series of modified methods assumed that the rate of nucleotide substitutions approximately follows the gamma distribution, and introduced the shape parameter a into conventional methods of calculating Ka and Ks. Therefore, these new methods can be regarded as the generalization of conventional approximate methods. An approximate method usually involves three steps (*1*, *2*):

1. Count synonymous and nonsynonymous sites;
2. Count synonymous and nonsynonymous differences;
3. Calculate the proportions of differences and correct for multiple hits.

We describe the modified methods step by step focusing on the modifications.

γ-NG method

γ-NG performs in the same mode as NG does in the procedures of counting sites and counting differences (*6*). Now we have

$$p_n = N_d/N \qquad (1)$$

$$p_s = S_d/S \qquad (2)$$

However, it uses a modified JC69 model to correct for multiple hits as follows (see more details in Supporting Online Material) (*32*):

$$\overline{d} = 3\,\overline{\alpha t} = \frac{3a}{4}\left[\left(1-\frac{4}{3}\overline{P}\right)^{-\frac{1}{a}}-1\right] \qquad (3)$$

As a result, we have

$$Ka = \frac{3a}{4}\left[\left(1-\frac{4}{3}p_n\right)^{-\frac{1}{a}}-1\right] \qquad (4)$$

$$Ks = \frac{3a}{4}\left[\left(1-\frac{4}{3}p_s\right)^{-\frac{1}{a}}-1\right] \qquad (5)$$

γ-LWL method

In comparison with LWL method (*7*), we pay more attention to the estimation for the number of transitional and transversional substitutions. We denote P_i and Q_i as the number of observed transitional and transversional differences at i-fold degenerate sites according to L_i (i=0, 2 or 4), which means the number of sites in the three corresponding degeneracy categories averaging over paired sequences. To compute the number of transitional (A_i) and transversional (B_i) substitutions per site (i=0, 2 or 4), we apply a modified K80 model based on P_i and Q_i as follows (see more details in Supporting Online Material) (*33*):

$$\begin{aligned} A_i &= \overline{\alpha t} \\ &= \frac{a}{2}\left[(1-2P_i-Q_i)^{-\frac{1}{a}}-1\right]-\frac{a}{4}\left[(1-2Q_i)^{-\frac{1}{a}}-1\right] \end{aligned} \qquad (6)$$

$$B_i = 2\,\overline{\beta t} = \frac{a}{2}\left[(1-2Q_i)^{-\frac{1}{a}}-1\right] \qquad (7)$$

And the subsequent procedures are the same as those in LWL method (*7*):

$$Ka = \frac{L_2B_2+L_0d_0}{2L_2/3+L_0} \qquad (8)$$

$$Ks = \frac{L_2A_2+L_4d_4}{L_2/3+L_4} \qquad (9)$$

where $d_i = A_i + B_i$ (i = 0, 2, or 4).

γ-LPB method

γ-LWL leaves out the transition/transversion rate difference in the procedure of counting two-fold site as 1/3 synonymous and 2/3 nonsynonymous, giving rise to underestimation of S and overestimation of Ks (and underestimation of Ka) and thus underestimation of ω (Ka/Ks). To overcome this drawback, we follow the same strategy as that in LPB method (*9*, *10*):

$$Ka = A_0 + \frac{L_2 B_2 + L_0 B_0}{L_2 + L_0} \tag{10}$$

$$Ks = \frac{L_2 A_2 + L_4 A_4}{L_2 + L_4} + B_4 \tag{11}$$

γ-MLWL method and γ-MLPB method

γ-MLWL follows another strategy to solve the problem that γ-LWL may perform poorly for large κ , as below (*8*):

When $\kappa \geq 2$,

$$Ka = \frac{L_2 B_2 + L_0 d_0}{\frac{2L_2}{(\kappa-1)+2} + L_0} \tag{12}$$

$$Ks = \frac{L_2 A_2 + L_4 d_4}{\frac{(\kappa-1)L_2}{(\kappa-1)+2} + L_4} \tag{13}$$

When $\kappa < 2$,

$$Ka = \frac{L_2 B_2 + L_0 d_0}{\frac{2L_2}{3} + L_0} \tag{14}$$

$$Ks = \frac{L_2 A_2 + L_4 d_4}{\frac{L_2}{3} + L_4} \tag{15}$$

where $d_i = A_i + B_i$ (i = 0, 2, or 4).

We also correct for arginines as described in the literature for complex conditions based on LWL method and LPB method (*8*) and denoted the modified versions as γ-MLWL and γ-MLPB.

γ-YN method

Our γ-YN method introduces gamma distribution into YN method (*5*), categorized with modified HKY85 (*34*) and F84 (*20*). Compared with YN method, the changed components are as follows:

The modified HKY85-F84 model is adopted to estimate κ on the basis of the nondegenerate and fourfold-degenerate sites (for more details see Supporting Online Material).

$$\overline{\kappa}_{F84} = \frac{\overline{(\kappa+1)\,\beta t} - \overline{\beta t}}{\overline{\beta t}} = \frac{\overline{(\kappa+1)\,\beta t}}{\overline{\beta t}} - 1 = \frac{\overline{h}}{\overline{i}} - 1 \tag{16}$$

where

$$\begin{aligned}\overline{h} &= \overline{(\kappa+1)\,\beta t} \\ &= a\left\{\left[1 - \frac{1}{2\left(g_T g_C/g_Y + g_A g_G/g_R\right)}\,\overline{P}\right.\right. \\ &\quad \left.\left. - \frac{g_T g_C g_R/g_Y + g_A g_G g_Y/g_R}{2\left(g_T g_C g_R + g_A g_G g_Y\right)}\,\overline{Q}\right]^{-\frac{1}{a}} - 1\right\}\end{aligned} \tag{17}$$

and

$$\overline{i} = \overline{\beta t} = a\left[\left(1 - \frac{1}{2g_Y g_R}\overline{Q}\right)^{-\frac{1}{a}} - 1\right] \tag{18}$$

$$\overline{\kappa}_{HKY85} = 1 + \frac{g_T g_C/g_Y + g_A g_G/g_R}{g_T g_C + g_A g_G}\,\overline{\kappa}_{F84} \tag{19}$$

where $\overline{P}$ and $\overline{Q}$ stand for the proportions of transitional and transversional differences for each synonymous and nonsynonymous site groups, respectively.

The modified F84 model is used to correct for multiple substitutions in terms of the divergent distance.

$$\overline{t} = [4g_T g_C(1+\overline{\kappa}_{F84}/g_Y)+4g_A g_G(1+\overline{\kappa}_{F84}/g_R)+4g_Y g_R]\times\overline{i} \tag{20}$$

where g_R=$g_A + g_G$ and g_Y=$g_T + g_C$.

Optimal index

To determine the optimal parameter a, we established an optimal index:

$$f = \sum_{1\leq i\leq 10}(C_i^{\text{estimated}} - C_i^{\text{expected}})^2 \tag{21}$$

In this expression, the values of i from 1 to 10 stand for the κ_{R} values increasing from 1 to 10, fixing κ_{Y}=3.75, as used in the analyses. $C_i^{\text{estimated}}$ denotes the estimated values of ω, Ka or Ks, and C_i^{expected} denotes the expected values of ω, Ka or Ks, when κ_{R}=i. This function measures the deviation from expected values, regardless if the deviation is positive or negative. We calculate the f value that corresponds to six different a values and choose the minimal f value as the optimal.

Comparative analysis based on computer simulation and real data testing

We employed the "*evolver*" Monte Carlo program, implemented in the PAML (Phylogenetic Analysis by Maximum Likelihood) package (*35*), to generate evolving protein-coding sequences based on specified substitution models. To reduce the influence of stochastic errors, we generate 2,000 pairs of sequences with 400 codons in each simulation. We choose appropriate ranges of related parameters for computer simulations, including codon frequencies, gamma distribution shape parameter (a), divergence time (t), two ratios of transitional rate between purines (κ_R) and between pyrimidines (κ_Y) to transversional rate, and selective pressure ω. In principle, we focus on the performance of various γ-methods at the optimal values of a. Moreover, we usually use rice codon frequencies as the defaults. And ω=0.3, 1, and 3 are used to represent negative selection, neutral mutation, and positive selection, respectively, and parameter t=0.6 is considered as a constant value except special occasions (*5*, *36*, *37*). In view of the unequal transitional substitutions, we often fix κ_Y to 3.75 and allow κ_R to vary from 1 to 10. To weigh the accuracies of Ka and Ks estimations, we computed the expected Ka and Ks values using the following equations (*8*):

$$Ka = \frac{(S+N)\times\omega\times t}{3\times(S+\omega\times N)} \tag{22}$$

$$Ks = \frac{(S+N)\times t}{3\times(S+\omega\times N)} \tag{23}$$

We formulate the error rate with a common definition:

$$\text{error rate} = \frac{\text{estimated value} - \text{expected value}}{\text{expected value}}\times 100\%$$

To examine the performance of our new methods in real data, 14,725 human-dog, 16,368 human-mouse, and 15,646 human-chimp orthologous gene pairs were collected from NCBI's HomoloGene database (build 61) (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/). After eliminating ambiguous data (extremes in sequence homology), 14,309 human-dog, 16,046 human-mouse, and 12,278 human-chimp gene pairs were used for further analysis. In consideration of decreasing the random errors, we pooled the three datasets into one dataset, which was used for comparing the methods evaluated in this study.

Acknowledgements

We thank Dr. Zhang Zhang (Department of Ecology and Evolutionary Biology, Yale University) and Miss Yanyang Zhi (Institute of Genetics and Developmental Biology, Chinese Academy of Sciences) for their sincere help on this manuscript. This work was supported by the National Basic Research Program of China (Grant No. 2006CB910404) awarded to JY.

Authors' contributions

DW conducted mathematical calculation, performed computational simulation, collected and analyzed the data, and drafted the manuscript. DW and SZ conceived and designed this study. FH, JZ, and SH contributed to data analysis. JY supervised the study and revised the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Nei, M. and Kumar, S. 2000. *Molecular Evolution and Phylogenetics.* Oxford University Press, New York, USA.
2. Yang, Z. 2006. *Computational Molecular Evolution.* Oxford University Press, New York, USA.
3. Zhang, Z. and Yu, J. 2006. Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. *Genomics Proteomics Bioinformatics* 4: 173-181.
4. Muse, S.V. 1996. Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* 13: 105-114.
5. Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17: 32-43.
6. Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3: 418-426.
7. 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.

8. 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.
9. Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 36: 96-99.
10. 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.
11. Zhang, Z., *et al.* 2006. Computing Ka and Ks with a consideration of unequal transitional substitutions. *BMC Evol. Biol.* 6: 44.
12. Bofkin, L. and Goldman, N. 2007. Variation in evolutionary processes at different codon positions. *Mol. Biol. Evol.* 24: 513-521.
13. Kocher, T.D. and Wilson, A.C. 1991. Sequence evolution of mitochondrial DNA in humans and chimpanzees: control region and a protein-coding region. In *Evolution of Life: Fossils, Molecules, and Culture* (eds. Osawa, S. and Honjo, T.), pp.391-413. Springer, Tokyo, Japan.
14. 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.
15. Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J. Mol. Evol.* 37: 613-623.
16. Wakeley, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* 11: 436-442.
17. Jin, L. and Nei, M. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* 7: 82-102.
18. Li, W.H., *et al.* 1990. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc. Natl. Acad. Sci. USA* 87: 6703-6707.
19. Yang, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* 10: 1396-1401.
20. Yang, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. *J. Mol. Evol.* 39: 306-314.
21. Yang, Z., *et al.* 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11: 316-324.
22. Wang, D.P., *et al.* 2009. Gamma-MYN: a new algorithm for estimating Ka and Ks with consideration of variable substitution rates. *Biol. Direct* 4: 20.
23. Zhang, Z., *et al.* 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4: 259-263.
24. Li, J., *et al.* 2009. Correlation between Ka/Ks and Ks is related to substitution model and evolutionary lineage. *J. Mol. Evol.* 68: 414-423.
25. Yang, Z., *et al.* 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155: 431-449.
26. Berglund, A.C., *et al.* 2005. Tertiary windowing to detect positive diversifying selection. *J. Mol. Evol.* 60: 499-504.
27. Fares, M.A. 2004. SWAPSC: sliding window analysis procedure to detect selective constraints. *Bioinformatics* 20: 2867-2868.
28. Fares, M.A., *et al.* 2002. A sliding window-based method to detect selective constraints in protein-coding genes and its application to RNA viruses. *J. Mol. Evol.* 55: 509-521.
29. Liang, H., *et al.* 2006. SWAKK: a web server for detecting positive selection in proteins using a sliding window substitution rate analysis. *Nucleic Acids Res.* 34: W382-384.
30. Siltberg, J. and Liberles, D.A. 2002. A simple covarion-based approach to analyse nucleotide substitution rates. *J. Evol. Biol.* 15: 588-594.
31. Suzuki, Y. 2004. Three-dimensional window analysis for detecting positive selection at structural regions of proteins. *Mol. Biol. Evol.* 21: 2352-2359.
32. Jukes, T.H. and Cantor, C.R. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. Munro, H.N.), vol.III, pp.21-132. Academic Press, New York, USA.
33. 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.
34. Hasegawa, M., *et al.* 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22: 160-174.
35. Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586-1591.
36. Li, W.H. 1997. *Molecular Evolution.* Sinauer Associates, Inc., Sunderland, USA.
37. Messier, W. and Stewart, C.B. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385: 151-154.

Supporting Online Material

Figures S1–S6, Tables S1 and S2, and other materials

DOI: 10.1016/S1672-0229(08)60040-6