J Mol Evol (2009) 68:414–423
DOI 10.1007/s00239-009-9222-9

Correlation Between Ka/Ks and Ks is Related to Substitution Model and Evolutionary Lineage

Jun Li · Zhang Zhang · Søren Vang · Jun Yu · Gane Ka-Shu Wong · Jun Wang

Received: 22 December 2008 / Accepted: 27 February 2009 / Published online: 24 March 2009

Abstract In 2005, Wyckoff and coworkers described a surprisingly strong correlation between Ka/Ks and Ks in several data sets using the LPB93 algorithm. This finding indicated the possibility of a paradigm shift in the way selection strength can be measured using the Ka/Ks ratio. We carried out a calculation of Ka and Ks using six different algorithms on three cross-species orthologous data sets and found a highly variable correlation among the algorithms and lineages. Algorithms based on the GY-HKY substitution model exhibit a weaker positive correlation or a stronger negative correlation than those based on the K2P and JC69 substitution model. Even if one algorithm shows a positive correlation between Ka/Ks and Ks in a warm-blooded lineage, it may show no correlation in a cold-blooded lineage. This algorithm-related and evolutionary lineage-related correlation indicates the need for great caution in drawing conclusions when using only one Ka and Ks algorithm in a genomewide analysis of

Jun Li, Zhang Zhang and Søren Vang contributed equally to this work.

J. Li · Z. Zhang · J. Yu · G. K.-S. Wong · J. Wang (✉)
Beijing Genomics Institute, Shenzhen, Building Complex,
BeiShan Industrial Zone, Yantian District, Shenzhen 518083,
China
e-mail: wangj@genomics.org.cn

J. Li
e-mail: junli@genomics.org.cn

J. Yu
e-mail: junyu@big.ac.cn

Z. Zhang · J. Yu
Beijing Institute of Genomics, Chinese Academy of Sciences,
Beijing 100080, China

Z. Zhang
Graduate School of Chinese Academy of Sciences,
Beijing 100039, China
e-mail: zhang.zhang@yale.edu

S. Vang
Research Unit for Molecular Medicine, Aarhus University
Hospital, 8200 Aarhus N, Denmark
e-mail: vang@ki.au.dk

S. Vang
Faculty of Health Sciences, University of Aarhus, 8200 Aarhus
N, Denmark

J. Yu · G. K.-S. Wong
Key Laboratory of Genomic Bioinformatics of Zhejiang
Province, Hangzhou Genomics Institute, James D. Watson
Institute of Genome Sciences of Zhejiang University,
Hangzhou 310007, China

G. K.-S. Wong (✉)
Department of Biological Sciences, University of Alberta,
Edmonton, AB T6G 2E9, Canada
e-mail: gane@ualberta.ca

G. K.-S. Wong
Department of Medicine, University of Alberta, Edmonton,
AB T6G 2E9, Canada

J. Wang
Department of Biochemistry and Molecular Biology,
University of Southern Denmark, 5230 Odense M, Denmark

J. Wang
Institute of Human Genetics, University of Aarhus,
8000 Aarhus C, Denmark

Present Address:
Z. Zhang
Department of Ecology and Evolutionary Biology,
Yale University, New Haven, CT 06520, USA

selection strength. Our results indicated that currently used algorithms for Ka and Ks calculations are flawed and need improvements.

Keywords Ka/Ks · Substitution model · Evolutionary lineage related

Introduction

The nonsynonymous substitution rate (Ka), the synonymous substitution rate (Ks), and their ratio (Ka/Ks; sometimes termed dN/dS) are commonly used to aid in understanding the direction of evolution and its selective strength in a coding sequence (Fay and Wu 2001; Kimura 1983; Li 1997; Nei and Kumar 2000; Ohta 1995; Yang and Bielawski 2000). Ka/Ks $>$ 1 indicates a positive selection, Ka/Ks $<$1 indicates a negative selection, and Ka/Ks $\approx$1 indicates a neutral evolution.

A recent study described a surprisingly strong positive correlation between Ka/Ks and Ks in several data sets using the LPB93 algorithm (Wyckoff et al. 2005). This finding indicated the possibility of a paradigm shift in the way selection strength can be measured using the Ka/Ks ratio. The authors proposed that the Ka/Ks value reflects not only selective strength but also neutral mutation rate. A later study (Liao and Zhang 2006) did not show a strong correlation between Ka/Ks and Ks within mammalian orthologues using PAML and suggests that the correlation might be sensitive to the method or substitution model used.

Algorithms for estimating Ka and Ks normally involve three steps: counting the number of synonymous and nonsynonymous sites, counting the numbers of synonymous and nonsynonymous substitutions, and correcting for multiple substitutions (Yang and Nielsen 2000). These algorithms adopt different substitution or mutation models based on different assumptions that take various sequence features into account: this gives rise to varied estimates of evolutionary distance (Muse 1996). Thus, the estimation of Ka and Ks is sensitive to the underlying assumptions or mutation models (Zhang and Jun 2006).

Table 1 provides details on the characteristics of several of these types of algorithms, how their authors

Table 1 Characteristics and application of commonly used KaKs algorithms

Method	Ti/Tv[a]		Codon/ nucleotide frequencies	Evaluation method		Applied to genome project
	Sites	Substitutions		Computer simulation	Real data	
NG86	No	No	No	Use nucleotide frequencies from pseudogenes and test for a specific case for purifying selection with Ka/Ks = 0.2	Globin genes (human β vs. rabbit β, human β vs. chicken β, human β vs. human α1)	*Mycobacterium avium, Tetrahymena thermophila, Plasmodium yoelii yoelii*
LWL85	No	Yes[a]	No	–	40 genes from mammals (human, rodents, and artiodactyls)	*Betaherpesvirinae 6B*
LPB93	Yes	Yes	No	–	14 pairs of mouse and rat genes, 45 genes from human and mouse	*Plasmodium vivax, Rickettsia*
GY94	Yes	Yes	Yes	–	All pairwise comparisons of mammalian α- and β-globin genes	*Pan troglodytes, Canis familiaris*
YN00	Yes	Yes	Yes	Consider effects of codon frequencies, transition/transversion rate ratio, divergence time, and sequence length, respectively	Concatenated sequences of the 12 protein-coding genes on the H-strand of the mitochondrial genome from human and orangutan	*Mus musculus, Rattus norvegicus*
MYN	Yes[b]	Yes[b]	Yes	Consider effects of codon frequencies, two ratios of transitional rates between purines and between pyrimidines over the transversional rate, divergence time, and sequence length, respectively	Concatenated sequences on three genome-wide orthologous data sets (human-mouse, human-dog, mouse-rat)	

[a] Ti/Tv (transition/transversion) has a stronger influence on the number of sites than do substitutions. In general, the number of substitutions is less than that of the total number of sites, which results in similar trends between NG86 and LWL85 (as shown in Fig. 1)

[b] MYN considers unequal transition/transversion rate ratios, stemming from the assumption of different transitional rates between purines and between pyrimidines

evaluated them, and to which genome projects they were applied. The NG86 algorithm (Nei and Gojobori 1986), based on the Jukes–Cantor model (JC69; Jukes and Cantor1969), assumes substitutions with equal frequency and considers different evolutionary pathways between pairwise sequences. In contrast, the LWL85 algorithm (Li et al. 1985) introduces nondegenerate, twofold degenerate, and fourfold degenerate sites to count sites and substitutions, which is based on the two-parameter model of Kimura (K2P; Kimura 1980). Although K2P is used for correction of multiple substitutions, the LWL85 algorithm allows for different rates between transitions and transversions only by counting substitutions and considers that twofold degenerate sites are one-third synonymous and two-thirds nonsynonymous. In comparison with the LWL85 algorithm, LPB93 (Li 1993) takes into account such bias by counting sites, and the differences between the LWL85 and the LPB93 algorithms mainly focus on their Ka and Ks formulas. The GY94 algorithm (Goldman and Yang 1994) is a maximum-likelihood method that adopts a codon-based model (GY-HKY) considering more features of DNA sequence evolution, e.g., transition/transversion rate bias and nucleotide/codon frequency bias. The YN00 algorithm (Yang and Nielsen 2000) is a simplified version of the GY94 algorithm (Hasegawa et al. 1985) and gives a close approximation of this more time-consuming maximum-likelihood method. The MYN algorithm (Zhang et al. 2006) is a modification of YN00 and adopts the Tamura/Nei (1993) model, which considers unequal transitional rates between purines and pyrimidines, as well as considering transversional rate and nucleotide (codon) frequencies. Beside these methods, Ina's (1995) method and the modified NG method (Zhang et al. 1998) are also frequently used. Ina's method does not partition sites according to site degeneracy. However, it takes into account the transition/transversion rate bias by counting synonymous and nonsynonymous sites in proportion to synonymous and nonsynonymous substitution rates. The modified NG method considers the transition/transversion rate bias and estimates the number of synonymous and nonsynonymous site with the K2P model.

Most of these algorithms were introduced and evaluated using either simulated or small-scale real data (Table 1) but, as yet, have not been evaluated in a large-scale, genome-wide evaluation of real data. In this report, we show that there is a highly variable correlation among the above six Ka and Ks algorithms in calculations on three complete orthologue data sets. Our results indicate that the correlation between Ka/Ks and Ks is affected not only by the algorithms used, but also by different evolutionary lineages of the DNA sequences analyzed.

Data and Methods

Orthologue Data and Alignment

To define the orthologue genes and the alignments of human-mouse and mouse-rat, we retrieved orthologous gene data from NCBI Homologueene database (ftp://ftp.ncbi.nih.gov/pub/HomoloGene/; version 44.1) and all sequences of the Refseq data from the NCBI genome (ftp://ftp.ncbi.nih.gov/genomes/). Refseq orthologues in a one-to-many or many-to-many relationship were excluded to avoid creating ambiguous orthologue pairs. A total of 15,065 and 14,198 orthologous pairs, respectively, were defined for human-mouse and mouse-rat; 15,743 fugu-tetraodon one-to-one orthologue relationships were defined using the InParanoid database (http://inparanoid.cgb.ki.se). Each pair of orthologous proteins was aligned using the blastp program in the NCBI BLAST2 package and their final nucleotide alignment for the Ka and Ks calculation was created according to the protein alignment.

Calculation of Ka, Ks, and Divergence

The NG86, LWL85, LPB93, GY94, YN00, and MYN algorithms, implemented in KaKs_Calculator (Zhang Zhang 2006), were used on all data sets to calculate Ka and Ks. For GY94 we used an F3x4 codon frequency model and default values for other parameters. Model weight was also calculated using KaKs_Calculator. Divergence (D) between orthologue pairs was calculated according to the proportion distance (p-distance) of each orthologue at the nucleotide level.

Computer Simulation Method

We used the simulation program evolver in PAML (Phylogenetic Analysis by Maximum Likelihood [Yang 2007]; available at: http://abacus.gene.ucl.ac.uk/software/paml.html) to get simulation data. All evolution parameters were extracted from human-mouse orthologue alignments in this study. Codon usage was extracted from human Refseq. Transition/transversion rate (average value, 3.820), Ka/Ks (average value, 0.182 and 0.136 for LPB93 and YN00, respectively), and substitution rate t (average value, 0.589 and 0.657 for LPB93 and YN00, respectively) were extracted from each pair of human-mouse orthologue pairs in this study. The average nucleotide frequencies are 0.257, 0.219, 0.260, and 0.264 for A, T, C, and G, respectively. The codon frequency used for simulation can be found at http://evolution.genomics.org.cn/dNdS_corre/human.codon-usage. Each simulated orthologue pair was assigned a series of parameters, including Ka/Ks, transition/transversion rate, and t, from its trained parameters and then evolved using the GY-HKY and K2P substitution model.

Orthologue Filtering Method

After calculating divergence (D), the orthologue pairs with the largest divergence (upper 5%) were also removed to prevent inaccurate or ambiguously-defined alignments. After cleaning, there were a total of 14,311, 13,488, and 14,954 orthologue pairs defined for human-mouse, mouse-rat, and fugu-tetraodon, respectively.

Statistical Methods

GSL (GNU Scientific Library, www.gnu.org/software/gsl/) was used for statistical analyses with standard C.

Results

The Algorithm and its Underlying Substitution Model Impact Ka/Ks, Ks, and the Correlation Between Ka/Ks and Ks

We first used the six algorithms (NG86, LWL85, LPB93, GY94, YN00, and MYN) to assess the correlation between Ka/Ks and Ks in three vertebrate cross-species orthologues (Fig. 1a–c). The data show that these analyses provide different degrees of correlation between Ka/Ks and Ks. After calculating *t*-values for each case with H^0: $r = 0$ (r is the correlation coefficient), we calculate *p*-values for each case under a *t* distribution (Bernstein 1999). To make the original distribution of Ka/Ks vs Ks easily observable, we randomly selected 2000 original points with the YN00 algorithm for the fugu-tetraodon lineage (Fig. 1d), which shows a distinct correlation when Ks is low.

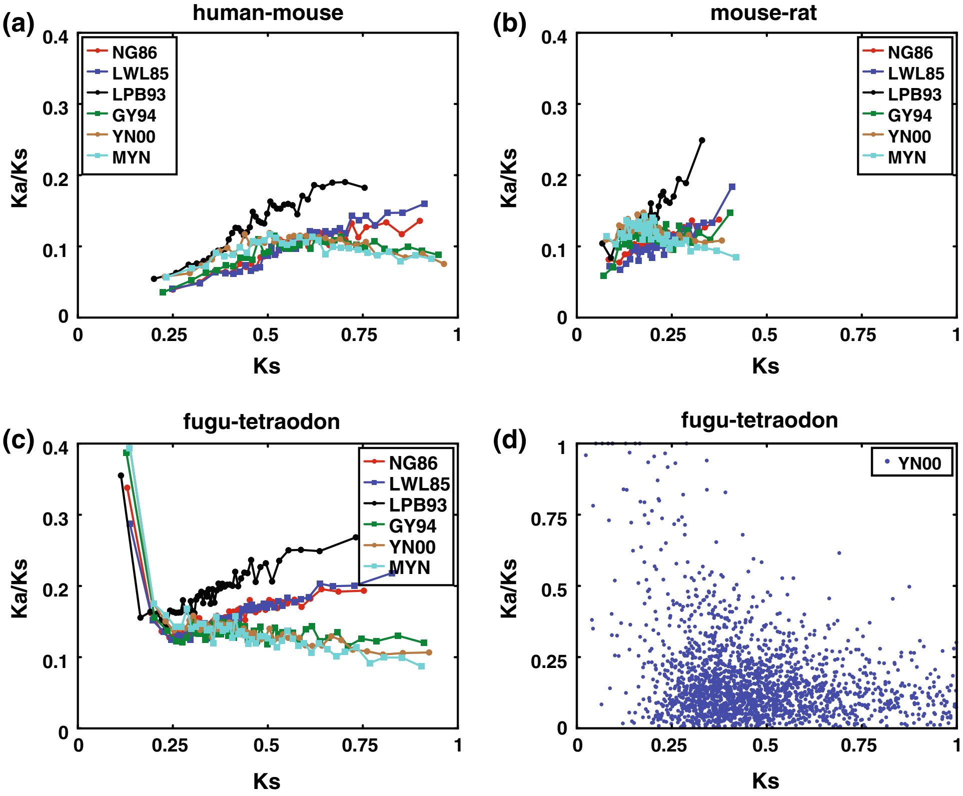

Fig. 1 Correlation between Ka/Ks and Ks for six different algorithms on three orthologue data sets. **a** Orthologues (15,065) of human-mouse with average Ks values of 0.610 (NG86), 0.606 (LWL85), 0.496 (LPB93), 0.748 (GY94), 0.755 (YN00), and 0.798 (MYN). Average Ka/Ks ratios are 0.136 (NG86), 0,140 (LWL85), 0.182 (LPB93), 0.131 (GY94), 0.136 (YN00), and 0.128 (MYN). **b** Orthologues (14,198) of mouse-rat with average Ks values of 0.223 (NG86), 0.231 (LWL85), 0.184 (LPB93), 0.217 (GY94), 0.219 (YN00), and 0.224 (MYN). Average Ka/Ks values are 0.155 (NG86), 0.151(LWL85), 0.202 (LPB93), 0.187 (GY94), 0.183 (YN00), and 0.176 (MYN). **c** Orthologues (15,743) of fugu-tetraodon with average Ks values of 0.433 (NG86), 0.439(LWL85), 0.379 (LPB93), 0.564 (GY94), 0.648 (YN00), and 0.624 (MYN). Average Ka/Ks values are 0.187 (NG86), 0.185 (LWL85), 0.224 (LPB93), 0.174 (GY94), 0.165 (YN00), and 0.158 (MYN). **d** Random selection of 2000 original points with the YN00 algorithm for the fugu-tetraodon lineage. All original Ka/Ks and Ks values for each orthologue pair were sorted by their Ks value in a, b, and c, and 300 consecutive points were placed in one bin. Subsequently, the mean values of Ka/Ks and Ks were calculated in each bin as representative Ka/Ks and Ks values for each bin

The NG86, LWL85, and LPB93 algorithms applied to the human-mouse and mouse-rat orthologues indicate there is a relatively strong positive correlation ($r^2 > 0.5$ and $p <$ 1e-7, for both human mouse and mouse-rat) between Ka/Ks and Ks (Table 2), whereas GY94 shows a much weaker correlation (human-mouse lineage, $r^2 = 0.28$ and $p =$ 5.88e-4; mouse-rat lineage, $r^2 = 0.035$ and $p =$ 1.24e-1). YN00 and MYN yield a weak negative correlation in human-mouse lineages ($r^2 < 0.3$ and $p >$ 1e-2) and a relatively strong negative correlation in mouse-rat lineages ($r^2 > 0.4$ and $p <$ 1e-6). For fugu-tetraodon lineages, NG86, LWL85, and LPB93 show almost no correlation between Ka/Ks and Ks ($r^2 < 0.05$ and $p > 0.1$). In contrast, GY94, YN00, and MYN exhibit a stronger negative correlation ($r^2 > 0.2$ and $p <$ 1e-3).

Figure 1 shows that the correlation between Ka/Ks and Ks is not consistent for all algorithms within a particular evolutionary lineage. Compared to the GY94, YN00, and MYN algorithms, NG86, LWL85, and LPB93 show a relatively strong positive correlation between Ka/Ks and Ks in human-mouse and mouse-rat lineages. In the fugu-tetraodon lineage, however, GY94, YN00, and MYN show a much weaker negative correlation than NG86, LWL85, and LPB93 do. The analyses here show that there is a general similarity in the correlation of NG86, LWL85, and LPB93, as there is also a similarity in the correlation of GY94, YN00, and MYN. The two groups, however, differ from each other.

If we consider the substitution models at the nucleotide level of these algorithms, we can group the six algorithms into three model groups (Posada and Crandall 2001). The first group, JC69 (Jukes and Cantor 1969), includes the NG86 algorithm. The second model group, K2P (Kimura 1980), includes the LWL85 and LPB93 algorithms. The third model group, GY-HKY (Goldman and Yang 1994), includes the GY94, YN00, and MYN algorithms.

To further test the substitution model's influence on the correlation between Ka/Ks and Ks, we carried out the following computer simulation to evaluate the correlation differences between group K2P (represented by LPB93) and group GY-HKY (represented by YN00). We did not assess JC69 because it contains an algorithm that performed similarly to those algorithms in K2P (Fig. 1; Table 2). The purpose of computer simulations is to examine whether different substitution models do or do not affect the correlation between Ka/Ks and Ks in simulation data.

Table 2 Correlation coefficient (*r*) and statistical significance for each algorithm under different evolutionary lineages

Method	Human-mouse			Mouse-rat			Fugu-tetraodon		
	r	r^2	*p*-value	*r*	r^2	*p*-value	*r*	r^2	*p*-value
NG86	0.937	0.878	<1e-17	0.758	0.575	1.73e-8	–0.136	0.018	1.71e-e-1
LWL85	0.962	0.925	<1e-17	0.846	0.716	6.22e-12	0.052	0.003	3.58e-e-1
LPB93	0.935	0.874	<1e-17	0.800	0.640	4.91e-10	0.032	0.001	4.12e-e-1
GY94	0.533	0.284	5.88e-e-4	0.187	0.035	1.24e-e-1	–0.464	0.215	3.40e-e-4
YN00	–0.545	0.297	2.95e-e-1	–0.706	0.498	3.75e-e-7	–0.560	0.314	1.81e-e-5
MYN	–0.356	0.127	2.47e-e-2	–0.768	0.590	9.06e-e-9	–0.610	0.372	3.47e-e-6

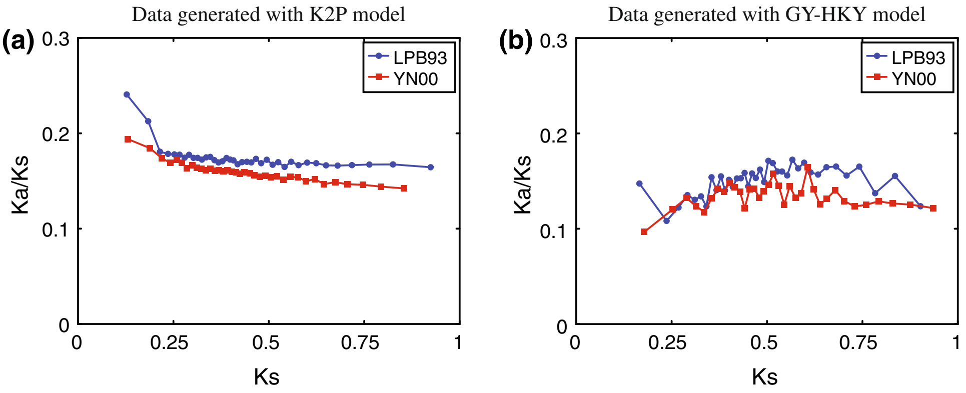

Fig. 2 Computer simulation for two substitution-model groups. K2P includes the LWL85 and LPB93 algorithms, and GY-HKY includes the GY94, YN00, and MYN algorithms. All simulated data were generated using **a** the K2P substitution model and **b** the GY-HKY substitution model. Ka and Ks calculations were carried out using both the LPB93 algorithm and the YN00 algorithm, and then the results were sorted by their Ks value, and 300 consecutive points were put in one bin. Subsequently, the mean value of Ka/Ks for each bin was used as a representative Ka/Ks for each bin

We first used the K2P substitution model to generate simulation data and then used LPB93 and YN00 to evaluate the correlation between Ka/Ks and Ks. The result (Fig. 2a) shows that YN00 provides a much stronger negative correlation than LPB93 does. The correlation coefficients for LPB93 and YN00 are –0.590 (p = 3.87e-5) and –0.917 (p = 7.77e-16), respectively. We then used the GY-HKY substitution model to generate our simulation sequences and calculated Ka and Ks using the LPB93 and YN00 algorithms (Fig. 2b). The difference in the correlation between the two algorithms is similar to the first simulation result. LPB93 shows a much stronger positive correlation than YN00. The correlation coefficients for LPB93 and YN00 are 0.401 (p = 6.3e-3) and 0.04 (p = 3.96e-1), respectively. Our simulation results confirm that data generated from LPB93 (the K2P model) will achieve a totally different correlation with YN00 (the GY-HKY model), and vice versa. For a detailed description of the simulation data, see Data and Methods. These results are supported by previous simulation studies (Tzeng et al. 2004), which show that when the evolutionary parameters are similar to those of the human-mouse lineage (CDS size, $\sim$ 400 codons; $\kappa \sim 2$; $t \sim 0.4$), Ks estimated by the GY-HKY model is larger than that estimated by the K2P model, and Ka/Ks estimated by the GY-HKY model is smaller than that by the K2P model (for details see Table 1 of Tzeng et al. 2004). Thus, the relative correlation between Ks and Ka/Ks for the same data set is very different and will result in different conclusions concerning selection direction and strength. We discuss the possible explanation of different degrees of correlation among these algorithms in detail in the Discussion.

Correlation Dependent on Evolutionary Lineage

We next compared the correlation between Ka/Ks and Ks in fixed algorithms for different evolutionary lineages Although NG86, LWL85, and LPB93 show a relatively strong positive correlation in human-mouse and mouse-rat lineages ($r^2 > 0.5$ and $p < 1\text{e-}7$), the correlation is lost in fugu-tetraodon ($r^2 < 0.05$ and $p > 0.1$) (Fig. 1; Table 2). GY94 shows a weak positive correlation in human-mouse (r^2 = 0.28 and p = 5.88e-4), no correlation in mouse-rat (r^2 = 0.035 and p = 1.24e-1), and a weak negative correlation in fugu-tetraodon (r^2 = 0.215 and p = 3.40e-4). YN00 and MYN show almost no correlation (or very weak negative correlation) in the human-mouse lineage ($r^2 < 0.3$ and $p > 1\text{e-}2$) but stronger negative correlations in mouse-rat ($r^2 > 0.4$ and $p < 1\text{e-}6$) and fugu-tetraodon ($r^2 > 0.3$ and $p < 1\text{e-}4$).

Figure 1 and Table 2 show that the correlation between Ka/Ks and Ks is related to lineages for a given algorithm, indicating that the correlation is also sensitive to the orthologue data. We used the following procedure to test the assumption: the correlation is related to the orthologue data for a specific algorithm. First, we calculated the

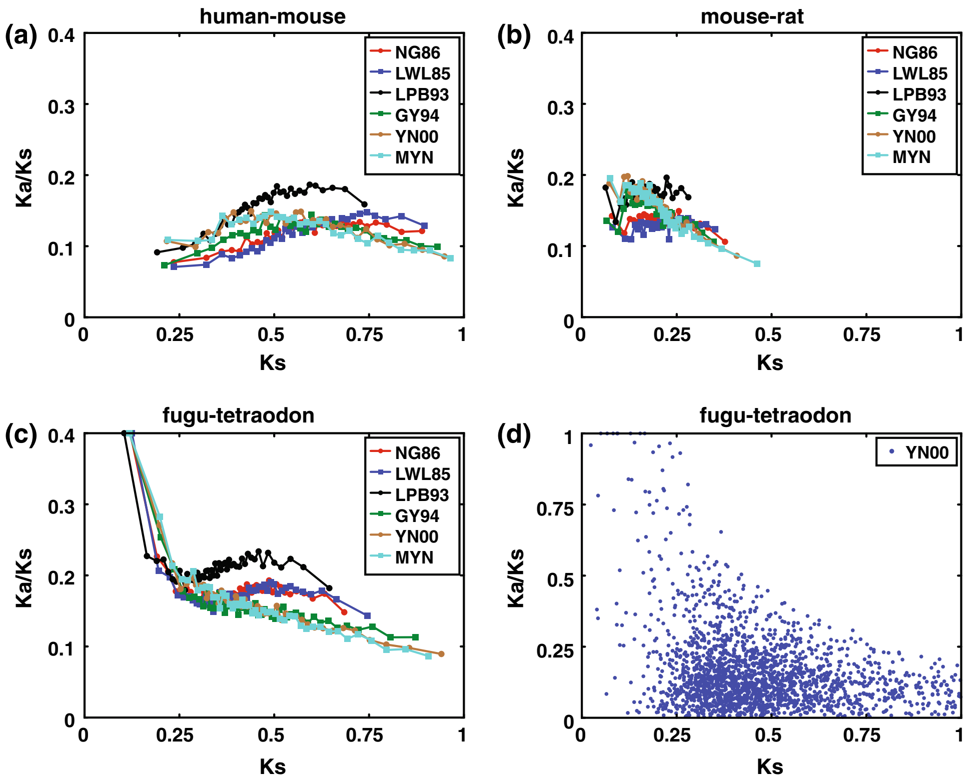

Fig. 3 Correlation between Ka/Ks and Ks for six different algorithms on three orthologue data sets after filtering the top 5% of divergent orthologues. **a** Orthologues (14,311) of human-mouse. **b** Orthologues (13,488) of mouse-rat. **c** Orthologues (14,954) of fugu-tetraodon. **d** Random selection of 2,000 original points using the YN00 algorithm for the fugu-tetraodon lineage. All original Ka/Ks and Ks values for each pair of orthologue pairs were sorted by their Ks value, and 300 consecutive points were put in one bin in **a**–**c**. Subsequently, the mean values of Ka/Ks and Ks were calculated for each bin as representative Ka/Ks and Ks values for each bin

Table 3 Correlation coefficient (r) and statistical significance for each algorithm under different evolutionary lineages after removing the top 5% of the divergent orthologue data

Method	Human-mouse			Mouse-rat			Fugu-tetraodon		
	r	r^2	p-value	r	r^2	p-value	r	r^2	p-value
NG86	0.753	0.567	3.81e-8	–0.437	0.191	3.40e-e-3	–0.399	0.159	2.24e-e-3
LWL85	0.901	0.812	5.92e-15	0.354	0.125	1.58e-2	–0.313	0.098	1.33e-2
LPB93	0.868	0.753	8.981e-13	0.264	0.070	5.69e-2	–0.249	0.062	4.06e-2
GY94	0.138	0.019	2.23e-1	–0.628	0.394	1.22e-5	–0.566	0.320	1.39e-5
YN00	–0.4418	0.195	5.08e-3	–0.931	0.867	<1e-17	–0.633	0.401	9.06e-7
MYN	–0.599	0.358	1.90e-4	–0.947	0.897	<1e-17	–0.661	0.436	3.94e-7

divergence level for each orthologous pair. Then we discarded the orthologues with the greatest divergence (top 5%) to avoid using incorrect and ambiguously defined alignments or orthologous relationships. For the remaining 95% of the orthologues, we recalculated the correlation coefficient and p-value for each algorithm.

Figure 3 and Table 3 show that, after this filtering procedure, almost all the algorithms present weaker positive correlations or stronger negative correlations than they do in the original orthologue data (in Fig. 1; Table 2). The changes in correlation using this procedure support the assumption that the correlation between Ka/Ks and Ks is also sensitive to evolutionary lineage and orthologue data. Correlation values for a particular algorithm can vary in different evolutionary lineages or in different data subsets (as shown in Figs. 1 and 3).

Alignment Quality Check

Table 4 presents data for the quality of the alignment in this study. All three evolutionary lineages were divided into two parts, according to their Ks value. For Ks $<$ 0.25, the percentages of mismatches in the alignments are 6.16%, 6.30%, and 7.15% for human-mouse, mouse-rat, and fugu-tetraodon, respectively. For Ks $>$ 0.25, the percentages of mismatches in the alignments are 14.99%, 9.93%, and 13.05% for human-mouse, mouse-rat, and fugu-tetraodon, respectively. The divergence for all three lineages is still in the region where Ka and Ks can be calculated accurately (Tzeng et al. 2004). We also assessed the gap rate for the whole alignments and the presence or absence of gaps at the ends of the alignments. Table 4 shows that for all three evolutionary lineages, the gap rates of both measurements are no more than 1%, except for human-mouse orthologues when Ks $<$ 0.25 (slightly more than 1%). This indicates that the alignment qualities are still reliable for all lineages.

To check whether the top 5% orthologues are really divergent from the other 95% of the orthologues, we further investigated the indel number, indel length, Ka, Ks, and Ka/Ks for the top 5% and the other 95% in all three evolutionary lineages. Table 5 shows that the indel number per gene and indel length per gene in the top 5% of the orthologues are about two to four times larger than those in

Table 4 Quality of alignment

	% mismatches	% gaps	% terminal gaps	GC content	GC3 content	Ts/Tv
Ks $<$ 0.25						
Human-mouse	6.157	0.064	1.010	0.485	0.482	6.087
Mouse-rat	6.304	0.058	0.000	0.514	0.571	5.014
Fugu-tetraodon	7.565	0.148	0.000	0.533	0.611	3.556
Ks $>$ 0.25						
Human-mouse	14.987	0.131	0.088	0.510	0.557	3.912
Mouse-rat	9.930	0.126	0.000	0.521	0.612	3.678
Fugu-tetraodon	13.048	0.220	0.081	0.527	0.631	2.453

Note. "% mismatches" indicates the average percentage of mutations in the alignment; lengths of gaps were excluded in the total length of the alignment. "% gaps" indicates the average percentage of gaps in all alignments and was calculated using the formula, % gaps = (total gap number/total alignment length) × 100. "% terminal gaps" indicates the percentage of orthologues that contain a gap at either end of their alignment and was calculated using the formula, % terminal gaps = (number of alignments that contain a gap at either end of the alignment/number of total alignments) × 100. Ts/Tv, which was measured with κHKY85 (Hasegawa et al. 1985), indicates the average transition/transversion rate ratio between mutations. GC and GC3 refer to GC content and GC content at the third position of all codons, respectively the rest of the orthlogues. Ka, Ks, and Ka/Ks in the top 5% of the divergent orthologues are significantly larger ($p < 0.005$, Wilcoxon rank sum test) than those in the rest of the orthologues. These results indicate that the top 5% of the divergent orthologues have more unreliable alignments than those of the other 95%.

In addition, to check whether the pattern observed in fugu-tetraodon was not an alignment artifact of our alignment procedure, we used another method for creating this alignment: global alignment with software “needle” based on the Needleman-Wunsch algorithm in EMBOSS (Rice et al. 2000). Using this method we obtained the same Ka/Ks vs. Ks pattern, and this appears at both low and high Ks values (data not shown). This concordance provides evidence to support the quality of our results, and that they are not related to the use of low-quality orthologue data or alignment artifacts.

Due to the potential influence of GC, GC3 content (GC content at the third position of all codons), and transition/transversion ratio on Ks estimation (Chamary et al. 2006), we examined the GC, GC3 content, and transition/transversion rate ratio (Ts/Tv) in each group of orthologues (Table 4). Our results show that when Ks > 0.25, the difference in GC (0.017) or GC3 (0.074) content between human-mouse and fugu-tetraodon orthologues is much smaller; in contrast, when Ks < 0.25, differences in GC (0.048) or GC3(0.129) content between human-mouse and fugu-tetraodon are larger. See the Discussion for more details about the influence of these parameters on Ks.

Discussion

Recently, Wyckoff and coworkers found a surprisingly strongly positive correlation between Ka/Ks and Ks using several data sets (Wyckoff et al. 2005) and suggested a paradigm shift in the application of Ka/Ks as a measure of selective strength, indicating that the Ka/Ks value reflects not only selective strength but also neutral mutation rate. In short lineages, the positive correlation between Ka/Ks and Ks is not observed (Wyckoff et al. 2005). However, after correcting the stochastic noise of Ks in short lineages, the positive correlation can still be observed with the LPB93 algorithm (Vallender and Lahn 2007). Although some earlier studies have shown some correlation between Ka and Ks (Domazet-Loso and Tautz 2003; Lynch and Conery 2000), so far, only Wyckoff et al. have presented a strong systematic correlation between Ka/Ks and Ks. In consideration of Wyckoff and coworkers’ findings, we analyzed three orthologue data sets with six different algorithms for evolutionary distance in three evolutionary lineages. Comparing NG86, LWL85, LPB93, GY94, YN00, and MYN, we found some correlations between Ka/Ks and Ks. However, those correlations had a highly variable strength and a dependence on the lineage used in these calculations.

Which factors might contribute to the cause of the phenomenon that different algorithms present different levels of correlation for the same data set? The following are possible interpretations: (1) transition/transversion rate bias, (2) codon usage bias, (3) the estimation difference among different substitution models increasing with increasing substitution rate, and (4) estimation error and imperfect computation for these KaKs algorithms. As discussed previously, the lack of incorporation of transition/transversion rate bias, NG86 will overestimate Ks and underestimate Ka/Ks (Yang 2006; Yang and Bielawski 2000). Ignoring codon-usage bias in NG86, LWL85, and LPB93 will result in underestimation of Ks and overestimation of Ka/Ks. When divergence increases, the estimation error will increase dramatically (Nei and Kumar 2000). That is, the percentage difference in Ks estimation between two substitution models will increase sharply

Table 5 Divergence between the top 5% and the other 95% of orthologues

	Human-mouse		Mouse-rat		Fugu-tetraodon	
	Top 5%	Other 95%	Top 5%	Other 95%	Top 5%	Other 95%
Mean no. of indels	7.9	2.0	4.8	1.0	9.7	3.1
Mean length of indels	114.8	27.0	102.4	26.1	199.0	72.0
GY94						
Mean Ka	0.404	0.077	0.195	0.031	0.293	0.069
Mean Ks	1.314	0.718	0.438	0.205	1.580	0.507
Mean Ka/Ks	0.378	0.118	0.622	0.164	0.283	0.167
LPB93						
Mean Ka	0.387	0.080	0.190	0.032	0.289	0.074
Mean Ks	0.821	0.478	0.340	0.175	0.788	0.356
Mean Ka/Ks	0.562	0.162	0.618	0.180	0.415	0.214

Note. According to the p-distance, all orthologues in the three lineages can be divided into two parts: top 5% and other 95%

when the synonymous substitution rate increases (Ks > 0.3). Because the nonsynonymous substitution rate is much lower than the synonymous substitution rate, the estimation difference between two substitution models is very small and will have little impact on the correlation. For comparison between LPB93 and YN00, when the synonymous substitution rate is low (<0.2), YN00 will show a little higher Ks and a little lower Ka/Ks than LPB93 does. When the synonymous substitution rate is higher (>0.3), YN00 will present a much higher Ks and much lower Ka/Ks than LPB93 (Yang 2006; Yang and Bielawski 2000). Therefore, assuming that LPB93 presents a straight line (strong positive correlation between Ka/Ks and Ks), YN00 will be like a parabola beneath the LPB93 straight line, which is consistent with our result in Figs. 1 and 3. Our simulation results further confirm that the systematic differences between the LPB93 and the GY94 algorithms exist for simulation data produced by either the K2P or the GY-HKY model. LPB93 will yield a more positive correlation between Ka/Ks and Ks, and GY94 will yield a more negative correlation for the same data set. One interesting result is that there is a small difference between GY94 and YN00: the negative correlation between Ka/Ks and Ks in YN is a little stronger than that in GY94, although they adopt the same underlying substitution model. This small difference can be explained by the numerical calculation difference between the maximum-likelihood and the approximate method. The approximate method (YN or MYN) will usually yield a little larger estimation of Ks and a little lower estimation of Ka/Ks (Yang and Bielawski 2000; Yang and Nielsen 2000), thus leading to a little more negative correlation than the maximum-likelihood method does (GY94). And these correlation differences between two very similar methods suggest that the impact of stochastic variance and imperfect computation on the correlation cannot be ignored. Additionally, the impact of stochastic noise on the correlation between Ka/Ks and Ks was also considered in recent studies (Vallender and Lahn 2007).

Why does a given KaKs algorithm lead to positive correlations for some data sets but negative correlations for other data sets? Two possible causes are as follows. (1) The change in “real” substitution will cause different degrees of correlation for different data sets even for the same algorithm. This interpretation can be confirmed by our simulation results in Fig. 2. (2) Evolutionary traits, such as codon frequency, transition/transversion rate bias, and divergence at the nucleotide level, will yield another conjunct impact on the correlation difference among different data sets. Different heterogeneity in the gene region, such as codon frequency, transition/transversion rate, CpG islands, and isochores (long stretches of compositionally homogeneous DNA), can also affect the Ka and Ks calculation performance. As for the comparison of warm-blooded and cold-blooded lineages, different compositional patterns of isochore structure exist. Additionally, cold-blooded vertebrate genomes have fewer GC regions and lack GC-rich isochores, which are widespread in warm-blooded vertebrate genomes. The GC3 content in a gene is highly correlated with the GC content of the isochore in which it is embedded in mammals (Chamary et al. 2006). Such a compositional difference in nucleotides may lead to codon usage bias and to different modes of genome evolution: conservative mode and transitional mode (Bernardi 1993). Our results also show that the difference in GC (or GC3) levels between human-mouse and fugu-tetraodon orthologues (we compared these two lineages due to their similar divergence) is much smaller when Ks > 0.25 than when Ks < 0.25: the correlation between Ka/Ks and Ks for human-mouse and fugu-tetraodon is different when Ks < 0.25 (negative correlation in fugu-tetraodon and positive correlation in human-mouse), whereas the correlation is similar between the two lineages when Ks > 0.25. This indicates that the isochore effect may result in different codon usage and selection bias on synonymous sites, as suggested previously (Chamary et al. 2006).

Although previous studies (Liao and Zhang 2006; Wyckoff et al. 2005) have employed correlation coefficients to measure the trend of Ka/Ks with Ks, our fugu-tetraodon results indicate that the trend or dependence of Ka/Ks on Ks is very complicated and very dependent on the regions (low or high Ks region) and calculation method used. Therefore, single statistics of measurement, such as correlation coefficient, may not provide a completely reliable or complete picture of the dependence of Ka/Ks on Ks. To ensure confidence in the current correlation assessment, it is necessary to avoid using global calculated correlations and also to consider the local dependence (correlation) of Ka/Ks and Ks.

Summary

Wyckoff and coworkers (2005) have shown a strong positive correlation between Ka/Ks and Ks with the LPB93 algorithm for human-mouse orthologues, which we have reproduced in this current study. However, when we calculated the correlation using several other algorithms (GY94, YN00, and MYN) and used more evolutionary lineages, including the cold-blooded fugu-tetraodon lineage, the positive correlation became less significant from warm-blooded to cold-blooded lineages using the NG86, LWL85, and LPB93 algorithms. At the same time, we found a weak or no significant negative correlation using GY94, YN00, and MYN in a warm-blooded lineage and stronger negative correlation using GY94, YN00, and

MYN in a cold-blooded lineage. In each evolutionary lineage, the correlation was variable among algorithms that are based on different DNA substitution models. Previously, algorithms to compute Ka and Ks were justified by how well they fit some arbitrarily defined mutation models. Given the algorithm-specific and evolutionary lineage-related correlations shown in this work, great caution should be taken when using only one Ka and Ks algorithm. A data set calculation with an improperly chosen algorithm may produce an inaccurate finding, which may then be interpreted as a biological trait but probably is an artifact of the calculation.

Acknowledgments We thank Laurie Goodman for her extensive editing of our manuscript. We thank Lin Fang, Jia Ye, and Hongkun Zheng for their constructive comments on this work. This project was supported by the Chinese Academy of Sciences (KSCX2-YW-N–023, GJHZ0701-6), the Ministry of Science and Technology under high-tech program 863 (2006AA02Z334, 2006AA10A121), the National Natural Science Foundation of China (90608010, 90403130, 30221004, 90612019, 30392130), and the Beijing Municipal Science and Technology Commission (D07030200740000). The project was also supported by Ole Rømer grants from the Danish Natural Science Research Council and the Danish Medical Research Council. J.L. collected the data and carried out all analyses. J.L., Z.Z., and S.V. wrote the manuscript. J.Y., G.K.W., and J.W. designed and coordinated the study. All authors read and approved the final manuscript.

References

Bernardi G (1993) The vertebrate genome: isochores and evolution. Mol Biol Evol 10:186–204

Bernstein SBR (1999) Elements of statistics II: inferential statistics. McGraw-Hill, New York

Chamary JV, Parmley JL, Hurst LD (2006) Hearing silence: non-neutral evolution at synonymous sites in mammals. Nat Rev Genet 7:98–108

Domazet-Loso T, Tautz D (2003) An evolutionary analysis of orphan genes in Drosophila. Genome Res 13:2213–2219

Fay JC, Wu CI (2001) The neutral theory in the genomic era. Curr Opin Genet Dev 11:642–646

Goldman N, Yang Z (1994) A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725–736

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Ina Y (1995) New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J Mol Evol 40:190–226

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HN (ed) Mammalian protein metabolism. Academic Press, New York, pp 21–123

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

Kimura M (1983) The neutral theory of molecular evolution. Cambridge University Press, Cambridge

Li WH (1993) Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 36:96–99

Li WH (1997) Molecular evolution. Sinauer Associates, Sunderland

Li WH, Wu CI, Luo CC (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

Liao B-Y, Zhang J (2006) Evolutionary conservation of expression profiles between human and mouse orthologous genes. Mol Biol Evol 23:530–540

Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. Science 290:1151–1155

Muse SV (1996) Estimating synonymous and nonsynonymous substitution rates. Mol Biol Evol 13:105–114

Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418–426

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. Oxford University Press, New York

Ohta T (1995) Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol 40: 56–63

Posada D, Crandall KA (2001) Selecting the best-fit model of nucleotide substitution. Syst Biol 50:580–601

Rice P, Longden I, Bleasby A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276–277

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Tzeng YH, Pan R, Li WH (2004) Comparison of three methods for estimating rates of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 21:2290–2298

Vallender E, Lahn B (2007) Uncovering the mutation-fixation correlation in short lineages. BMC Evol Biol 7:168

Wyckoff GJ, Malcom CM, Vallender EJ, Lahn BT (2005) A highly unexpected strong correlation between fixation probability of nonsynonymous mutations and mutation rate. Trends Genet 21: 381–385

Yang Z (2006) Computational molecular evolution. Oxford University Press, Oxford

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591

Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15:496–503

Yang Z, Nielsen R (2000) Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 17:32–43

Zhang J, Rosenberg HF, Nei M (1998) Positive Darwinian selection after gene duplication in primate ribonuclease genes. Proc Natl Acad Sci USA 95:3708–3713

Zhang Z, Jun Y (2006) Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. Genom Proteom Bioinform 4:173–181

Zhang Z, Li J, Yu J (2006) Computing Ka and Ks with a consideration of unequal transitional substitutions. BMC Evol Biol 6:44

Zhang Z, Li J, Zhao X, Wang J, Wong GK, Yu J (2006) KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genom Proteo Bioinform 4(4):259–263