KaKs_Calculator: Calculating Ka and Ks Through Model Selection and Model Averaging

Zhang Zhang[1,2,3#], Jun Li[2#], Xiao-Qian Zhao[2,3], Jun Wang[1,2,4], Gane Ka-Shu Wong[2,4,5], 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 of Chinese Academy of Sciences, Beijing 100049, China; [4] James D. Watson Institute of Genome Sciences, Zhejiang University, Hangzhou 310007, China; [5] UW Genome Center, University of Washington, Seattle, WA 98195, USA.

KaKs_Calculator is a software package that calculates nonsynonymous (Ka) and synonymous (Ks) substitution rates through model selection and model averaging. Since existing methods for this estimation adopt their specific mutation (substitution) models that consider different evolutionary features, leading to diverse estimates, KaKs_Calculator implements a set of candidate models in a maximum likelihood framework and adopts the Akaike information criterion to measure fitness between models and data, aiming to include as many features as needed for accurately capturing evolutionary information in protein-coding sequences. In addition, several existing methods for calculating Ka and Ks are also incorporated into this software. KaKs_Calculator, including source codes, compiled executables, and documentation, is freely available for academic use at http://evolution.genomics.org.cn/software.htm.

Key words: model selection, model averaging, AIC, approximate method, maximum likelihood method

Introduction

Calculating nonsynonymous (Ka) and synonymous (Ks) substitution rates is of great significance in reconstructing phylogeny and understanding evolutionary dynamics of protein-coding sequences across closely related and yet diverged species (*1–3*). It is known that Ka and Ks, or often their ratio (Ka/Ks), indicate neutral mutation when Ka equals to Ks, negative (purifying) selection when Ka is less than Ks, and positive (diversifying) selection when Ka exceeds Ks. Therefore, statistics of the two variables in genes from different evolutionary lineages provides a powerful tool for quantifying molecular evolution.

Over the past two decades, several methods have been developed for this purpose, which can generally be classified into two classes: approximate method and maximum likelihood method. The approximate method involves three basic steps: (1) counting the numbers of synonymous and nonsynonymous sites, (2) calculating the numbers of synonymous and nonsynonymous substitutions, and (3) correcting for multiple substitutions. On the other hand, the maximum likelihood method integrates evolutionary features (reflected in nucleotide models) into codon-based models and uses the probability theory to finish all the three steps in one go (*4*). However, these methods adopt different substitution or mutation models based on different assumptions that take account of various sequence features, giving rise to varied estimates of evolutionary distance (*5*). In other words, Ka and Ks estimation is sensitive to underlying assumptions or mutation models (*3*). In addition, since the amount and the degree of sequence substitutions vary among datasets from diverse taxa, a single model or method may not be adequate for accurate Ka and Ks calculations. Therefore, a model selection step, that is, to choose a best-fit model when estimating Ka and Ks, becomes critical for capturing appropriate evolutionary information (*6*, *7*).

Toward this end, we have applied model selection and model averaging techniques for Ka and Ks estimations. We use a maximum likelihood method based on a set of candidate substitution models and adopt the Akaike information criterion (AIC) to measure fitness between models and data. After choosing the

[#]Contributed equally to this work.
***Corresponding author.**
E-mail: junyu@genomics.org.cn

best-fit model for calculating Ka and Ks, we average the parameters across the candidate models to include as many features as needed since the true model is seldom one of the candidate models in practice (*8*). Finally, these considerations are incorporated into a software package, namely KaKs_Calculator.

Algorithm

Candidate models

Substitution models play a significant role in phylogenetic and evolutionary analyses of protein-coding sequences by integrating diverse processes of sequence evolution through various assumptions and providing approximations to datasets. We focused on a set of time-reversible substitution models (*9–16*) as shown in Table 1 (*17*, *18*), ranging from the Jukes-Cantor (JC) model, which assumes that all substitutions have equal rates and equal nucleotide frequencies, to the general time-reversible (GTR) model that considers six different substitution rates and unequal nucleotide frequencies. Subsequently, we incorporated the parameters in each nucleotide model into a codon-based model (*19*, *20*). As a result, a general formula of the substitution rate q_{ij} from any sense codon i to j ($i \neq j$) is given for all candidate models (*19*):

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ and } j \text{ differ by more than one difference} \\ \kappa_{xy}\,\pi_j & \text{if } i \text{ and } j \text{ differ by a synonymous substitution of } x \text{ for } y \\ \omega\,\kappa_{xy}\,\pi_j & \text{if } i \text{ and } j \text{ differ by a nonsynonymous substitution of } x \text{ for } y \end{cases}$$

where π_j is the frequency of codon j, ω is the Ka/Ks ratio, and κ_{xy} is the ratio of r_{xy} to r_{CA}, $x, y \in$ {A, C, G, T} (Table 1). For example, in the JC model, κ_{xy} and π_j are equal to 1 owing to equal substitution rates and equal nucleotide frequencies assumed. In the Hasegawa-Kishino-Yano (HKY) model, κ_{TC} and κ_{AG} become equivalent to the transition/transversion rate ratio and π_j can be estimated from sequences, similar to the method reported by Goldman and Yang (*19*). Other models can be accommodated by making obvious modifications. Therefore, we could acquire maximum likelihood scores in various values generated from individual candidate model by implementing the codon-based models in a maximum likelihood framework (*19*, *20*).

Model selection

AIC (*21*) has been widely used in model selection aside from other methods such as the likelihood ratio test (LRT) and the Bayesian information criterion (BIC) (*8*). AIC characterizes the Kullback-Leibler distance between a true model and an examined model, and this distance can be regarded as quantifying the information lost by approximating the true model. KaKs_Calculator uses a modification of AIC (AIC_C), which takes account of sampling size (n), maximum likelihood score ($\ln L_i$), and the number of parameters (k_i) in model i as follows:

$$AIC_{Ci} = AIC_i + \frac{2k_i(k_i+1)}{n-k_i-1} = -2\ln L_i + 2k_i + \frac{2k_i(k_i+1)}{n-k_i-1}$$

AIC_C is proposed to correct for small sampling size, and it approaches to AIC when sampling size comes to infinity. Consequently, we could use this equation to compute AIC_C for each candidate model and then identify a model that possesses the smallest AIC_C, which is a sign for appropriateness between models and data.

Model averaging

Model selection is merely an approximate fit to a dataset, whereas a true evolutionary model is seldom one of the candidate models (*8*). Therefore, an alternative way is model averaging, which assigns each candidate model a weight value and engages more than one model to estimate average parameters across models. Accordingly, we first need to compute the Akaike weight (w_i, where $i = 1, 2, \ldots, m$) for each model in a set of candidate models:

$$w_i = \frac{\exp[-\frac{1}{2}(AIC_{Ci} - \min AIC_C)]}{\sum_{j=1}^{m} \exp[-\frac{1}{2}(AIC_{Cj} - \min AIC_C)]}$$

where $\min AIC_C$ is the smallest AIC_C value among candidate models. We can then estimate model-averaged parameters. Taking κ_{TC} as an example, a model-averaged estimate can be calculated by:

$$\kappa_{TC} = \frac{\sum_{i=1}^{m} [w_i \times I(\kappa_{TC,i}) \times \kappa_{TC,i}]}{\sum_{i=1}^{m} [w_i \times I(\kappa_{TC,i})]}$$

where $\kappa_{TC,i}$ is κ_{TC} in model i and

$$I(\kappa_{TC,i}) = \begin{cases} 1 & \text{if } r_{TC} \neq r_{CA} \text{ in model } i \\ 0 & \text{otherwise} \end{cases}$$

Table 1 Candidate Models for Model Selection and Model Averaging in KaKs_Calculator

Model	Description (Reference)	Nucleotide frequency	Substitution rate*
JC	Jukes-Cantor model (*9*)	Equal	$r_{TC} = r_{AG} = r_{TA} = r_{CG} = r_{TG} = r_{CA}$
F81	Felsenstein's model (*10*)	Unequal	
K2P	Kimura's two-parameter model (*11*)	Equal	$r_{TC} = r_{AG} \neq r_{TA} = r_{CG} = r_{TG} = r_{CA}$
HKY	Hasegawa-Kishino-Yano model (*12*)	Unequal	
TNEF	TN model with equal nucleotide frequencies	Equal	$r_{TC} \neq r_{AG} \neq r_{TA} = r_{CG} = r_{TG} = r_{CA}$
TN	Tamura-Nei model (*13*)	Unequal	
K3P	Kimura's three-parameter model (*14*)	Equal	$r_{TC} = r_{AG} \neq r_{TA} = r_{CG} \neq r_{TG} = r_{CA}$
K3PUF	K3P model with unequal nucleotide frequencies	Unequal	
TIMEF	Transition model with equal nucleotide frequencies	Equal	$r_{TC} \neq r_{AG} \neq r_{TA} = r_{CG} \neq r_{TG} = r_{CA}$
TIM	Transition model	Unequal	
TVMEF	Transversion model with equal nucleotide frequencies	Equal	$r_{TC} = r_{AG} \neq r_{TA} \neq r_{CG} \neq r_{TG} \neq r_{CA}$
TVM	Transversion model	Unequal	
SYM	Symmetrical model (*15*)	Equal	$r_{TC} \neq r_{AG} \neq r_{TA} \neq r_{CG} \neq r_{TG} \neq r_{CA}$
GTR	General time-reversible model (*16*)	Unequal	

*r_{ij} indicates the rate of substitution of i for j, where $i, j \in \{A, C, G, T\}$.

Application

KaKs_Calculator is written in standard C++ language. It is readily compiled and run on Unix/Linux or workstation (tested on AIX/IRIX/Solaris). In addition, we use Visual C++ 6.0 for graphic user interface and provide its Windows version that can run on any IBM compatible computer under Windows operating system (tested on Windows 2000/XP). Compiled executables on AIX/IRIX/Solaris and setup application on Windows, as well as source codes, example data, instructions for installation and documentation for KaKs_Calculator is available at http://evolution.genomics.org.cn/software.htm.

Different from other existing tools (*22*, *23*), KaKs_Calculator employs model-selected and model-averaged methods based on a set of candidate models to estimate Ka and Ks. It integrates as many features as needed from sequence data and in most cases gives rise to more reliable evolutionary information (see the comparative results on simulated sequences at http://evolution.genomics.org.cn/doc/SimulatedResults.xls) (*24*). KaKs_Calculator also provides comprehensive information estimated from compared sequences, including the numbers of synonymous and nonsynonymous sites and substitutions, GC contents, maximum likelihood scores, and AIC_C. Moreover, KaKs_Calculator incorporates several other methods (*19, 25-31*) and allows users to choose one or more methods at one running time (Table 2).

Table 2 Methods Incorporated in KaKs_Calculator

Approximate method				
Method	Mutation model[#1]			Reference
	Step 1	Step 2	Step 3	
NG	JC	JC	JC	*26*
LWL	JC	K2P	K2P	*28*
MLWL	K2P	K2P	K2P	*30*
LPB	–*	–*	K2P	*25*, *29*
MLPB	–*	–*	K2P	*30*
YN	HKY	HKY	HKY	*27*
MYN	TN	TN	TN	*31*

Maximum likelihood method		
Method	Mutation model[#2]	Reference
GY	HKY	*19*, *20*
MS	a model that has the smallest AIC_C among 14 candidate models	Model-selected method proposed in this study
MA	a model that averages parameters across 14 candidate models	Model-averaged method proposed in this study

[#1]The approximate method involves three basic steps: Step 1: counting the numbers of synonymous and nonsynonymous sites; Step 2: calculating the numbers of synonymous and nonsynonymous substitutions; Step 3: correcting for multiple substitutions. [#2]The maximum likelihood method uses the probability theory to finish the three steps in one go (*4*). *No specific definition of synonymous and nonsynonymous sites or substitutions.

Although there exist 203 time-reversible models of nucleotide substitution (*8*), model selection in practice is often limited to a subset of them (*32*), and thus model averaging can reduce biases arising from model selection. Therefore, model-averaged methods should be preferred for general calculations of Ka and Ks. Some planned improvements include application of model selection and model averaging to detect positive selection at single amino acid sites, which requires high-speed computing for maximum likelihood estimation, especially when an adopted model becomes complex.

In conclusion, KaKs_Calculator incorporates as many features as needed for accurately extracting evolutionary information through model selection and model averaging, therefore it may be useful for in-depth studies on phylogeny and molecular evolution.

Acknowledgements

We thank Professor Ziheng Yang for the permission to use his invaluable source codes in PAML and two anonymous reviewers for their constructive comments on an earlier version of this manuscript. We are grateful to Ya-Feng Hu, Lin Fang, Jia Ye, Hai-Feng Yuan, and Heng Li for their help in software development. We also thank a number of users and members of our institutes for reporting bugs and giving suggestions. This work was supported by grants from the Ministry of Science and Technology of China (No. 2001AA231061) and the National Natural Science Foundation of China (No. 30270748) awarded to JY.

Authors' contributions

ZZ designed and programmed this software, and drafted the manuscript. JL carried out computer simulations to generate sequences. XQZ performed test for earlier versions of the software. JW and GKSW contributed in conceiving this software and participated in software design. 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. Kimura, M. 1983. *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK.
2. Li, W.H. 1997. *Molecular Evolution.* Sinauer Associates, Sunderland, USA.
3. Fay, J.C. and Wu, C.I. 2003. Sequence divergence, functional constraint, and selection in protein evolution. *Annu. Rev. Genomics Hum. Genet.* 4: 213-235.
4. Yang, Z. and Bielawski, J.P. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15: 496-503.
5. Muse, S.V. 1996. Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* 13: 105-114.
6. Sullivan, J. and Joyce, P. 2005. Model selection in phylogenetics. *Annu. Rev. Ecol. Evol. Syst.* 36: 445-466.
7. Pybus, O.G. 2006. Model selection and the molecular clock. *PLoS Biol.* 4: e151.
8. 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.
9. 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.
10. Felsenstein, J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 17: 368-376.
11. 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.
12. Hasegawa, M., *et al.* 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 22: 160-174.
13. 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.
14. Kimura, M. 1981. Estimation of evolutionary distances between homologous nucleotide sequences. *Proc. Natl. Acad. Sci. USA* 78: 454-458.
15. Zharkikh, A. 1994. Estimation of evolutionary distances between nucleotide sequences. *J. Mol. Evol.* 39: 315-329.
16. Tavare, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. *Lect. Math. Life Sci.* 17: 57-86.
17. Posada, D. 2003. Using Modeltest and PAUP* to select a model of nucleotide substitution. In *Current Protocols in Bioinformatics* (eds. Baxevanis, A.D., *et*

al.). John Wiley & Sons, New York, USA.
18. Lio, P. and Goldman, N. 1998. Models of molecular evolution and phylogeny. *Genome Res.* 8: 1233-1244.
19. Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11: 725-736.
20. 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.
21. Akaike, H. 1974. A new look at the statistical model identification. *IEEE Trans. Autom. Control* 19: 716-723.
22. Comeron, J.M. 1999. K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. *Bioinformatics* 15: 763-764.
23. Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13: 555-556.
24. Zhang, Z. and Yu, J. 2006. Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. *Genomics Proteomics Bioinformatics* 4: 173-181.
25. Li, W.H. 1993. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. *J. Mol. Evol.* 36: 96-99.
26. Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* 3: 418-426.
27. Yang, Z. and Nielsen, R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17: 32-43.
28. 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.
29. 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.
30. 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.
31. Zhang, Z., *et al.* 2006. Computing Ka and Ks with a consideration of unequal transitional substitutions. *BMC Evol. Biol.* 6: 44.
32. Posada, D. and Crandall, K.A. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics* 14: 817-818.