Available online at www.sciencedirect.com
ScienceDirect
www.sciencedirect.com/science/journal/16720229
GENOMICS PROTEOMICS & BIOINFORMATICS

Application Note

KaKs_Calculator 2.0: A Toolkit Incorporating Gamma-Series Methods and Sliding Window Strategies

Dapeng Wang[1,3#], Yubin Zhang[1,2,3#], Zhang Zhang[4], Jiang Zhu[1,3], and Jun Yu[1,2*]

[1]*CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China;*
[2]*Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China;*
[3]*Graduate University of Chinese Academy of Sciences, Beijing 100049, China;*
[4]*Plant Stress Genomics Research Center, Division of Chemical and Life Sciences & Engineering, King Abdullah University of Science and Technology, Thuwal 23955-6900, Saudi Arabia.*

Genomics Proteomics Bioinformatics 2010 Mar; 8(1): 77-80. DOI: 10.1016/S1672-0229(10)60008-3

Abstract

We present an integrated stand-alone software package named KaKs_Calculator 2.0 as an updated version. It incorporates 17 methods for the calculation of nonsynonymous and synonymous substitution rates; among them, we added our modified versions of several widely used methods as the gamma series including γ-NG, γ-LWL, γ-MLWL, γ-LPB, γ-MLPB, γ-YN and γ-MYN, which have been demonstrated to perform better under certain conditions than their original forms and are not implemented in the previous version. The package is readily used for the identification of positively selected sites based on a sliding window across the sequences of interests in 5′ to 3′ direction of protein-coding sequences, and have improved the overall performance on sequence analysis for evolution studies. A toolbox, including C++ and Java source code and executable files on both Windows and Linux platforms together with a user instruction, is downloadable from the website for academic purpose at https://sourceforge.net/projects/kakscalculator2/.

Key words: Ka/Ks, gamma-series methods, sliding window, positively selected sites

Introduction

Calculating nonsynonymous (Ka) and synonymous (Ks) substitution rates is a useful way for evaluating sequence variations for protein orthologs across different species or taxonomical lineages with unknown evolutionary status. Furthermore, it is often important to recognize positively selected sites and to identify genes with selective hotspots. There have been numerous methods and software tools developed for such purposes in the public domain, including PAML (*1*), MEGA (*2*), DnaSP (*3*), HyPhy (*4*) and certain modules from Bioperl (*5*). However, after careful simulations and real data analysis, we believe that a single method will not be readily identified to be used under all circumstances (*6, 7*), therefore we created the version of the KaKs_Calculator 1.0 (*8*), which adopted model-selected and model-averaged techniques to compute Ka/Ks values by means of a group of existing nucleotide substitution models.

Since the majority of DNA sequence sites are considered to be invariable due to functional restraints and evolutionary distances, the selective pressure varies among different sites in a sequence, thus Ka/Ks

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

calculations only based on the entire gene are not enough to detect the individual sites subjected to adaptive selection. To conquer this problem, a "sliding window" strategy has been introduced to several web servers such as SWAPSC (*9*) and WSPMaker (*10*), while these tools adopted fewer (mostly one) models for Ka and Ks calculations. Here we provide an updated version of KaKs_Calculator, which solves these two questions in a simple way. In particular, we have embedded gamma-series methods into this new version.

New Features

We have brought up three novel features into KaKs_Calculator 2.0. First, unlike the existent Ka/Ks algorithms, the new software can take the variable mutation rates across sequence sites into account, which contain vital information for molecular evolutionary studies. We created seven related methods namely γ-NG, γ-LWL, γ-MLWL, γ-LPB, γ-MLPB, γ-YN and γ-MYN by introducing gamma distribution to model the mutation rates; the importance of the new methods has been demonstrated as the ignorance gives rise to biased computational results (*11, 12*). We therefore implemented these new methods into the updated core tool of version 2.0, whose core toolset has seventeen algorithms including seven original approximate methods, seven gamma-series methods, one maximum likelihood method (GY), and two expanding methods (model selected and model averaged). The methods provide not only the values of Ka, Ks and Ka/Ks, but also other key information from paired orthologous sequences, including the number of synonymous/nonsynonymous sites, substitutions, divergence time, substitution-rate-ratio, GC content, and AICc. Second, we added three new modules—*split*, *plot*, *dpss*—to evaluate adaptive selection at the gene sequence level. As an expanding toolset, they adopt a sliding window with user's definition on window length and step length. *Split* is responsible for the division of the raw paired orthologous sequences into portions on the basis of dynamic windows in the positive direction. *Plot* deals with the outcome of the core toolset after the nucleotide sequences from *Split* have been computed, resulting in a massive collection of figures illustrating Ka, Ks and Ka/Ks (omega) in intervals. *Dpss* identifies the positions of positively selected sites based on the initial analyses. Third, it should be emphasized that all above-mentioned processes are capable of handling massive data in a timely fashion. In particular, all transferrable data including sequences and resulting information are contained in a single file. We provide executable files as well as source codes for the package and tested all programs on both Windows 2000/XP/Vista and Linux (Red Hat 3.4.6-8) platforms. The toolkit is freely available (licensed under GPLv3) online at https://sourceforge.net/projects/kakscalculator2/.

Implementation

In order to conveniently update the algorithm and to friendly communicate with users, we implement the new version with a "toolkit" idea in mind. Therefore, the integrated software is divided into two essential parts to better serve for different functionality: the core toolset that calculates Ka and Ks, and the expanding toolset that is responsible for additional computation activities based on the Ka and Ks calculation (*e.g.*, with a sliding window strategy) (**Figure 1**). In the core toolset, we design the GUI with visual C++'s MFC (Microsoft Foundation Classes) that manages documents and allows users to view the objects, and the entire program is object-oriented. Each main method has its own class in the code and the multi-thread operations among them use the CPU time allocations very efficiently. We adopt Java-6 to program the expanding toolset because of its advantages across different platforms. We choose R language (http://www.r-project.org/) to draw high-level graphics from inputting data. To call for the R function from Java, we employ a package named "Rserve" (http://www.rforge.net/Rserve/index.html), which is a program responding to requests from clients based on the TCP/IP protocol. In details, we use java to invoke the JRclient suite and connect it after Rserve starts on R environment; under this circumstance each connection has its workspace and directory. Moreover, the server allows many clients to plot their data simultaneously. In consideration of the running speed, it is so fast that a graph covering thousands of data points can be plotted in a few seconds.

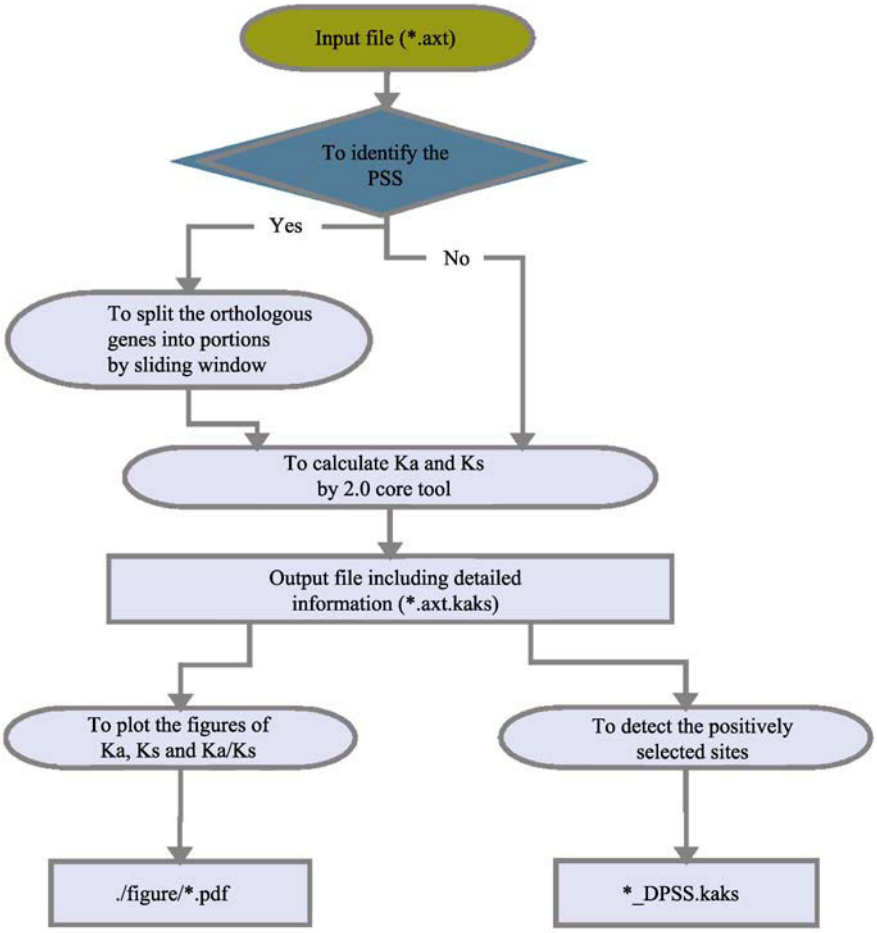

Figure 1 A flowchart of software design on KaKs_Calculator 2.0.

Evaluation

We have evaluated the performance of the gamma-series methods in Ka/Ks calculations in previous studies (*11, 12*). In the process of identifying positively selected sites, we have also successfully applied the toolbox to two real cases, including the animal alpha-defensin genes investigated in Lynn *et al* (*13*) and the *TAS1R3* (taste receptor type 1 member 3) genes reported to be responsible for the ability to recognize the sweetness (*14*) (**Figure 2**). It is important to combine the gamma-series methods with a sliding window strategy; the former represents the variation of raw mutation across sites and the latter reveals if each site is driven by different selective pressure based on the assumption that the omega (Ka/Ks) values are not equal across orthologous gene sequences. In particular, when window slices become dense enough, it approaches the "site models" (*1*),

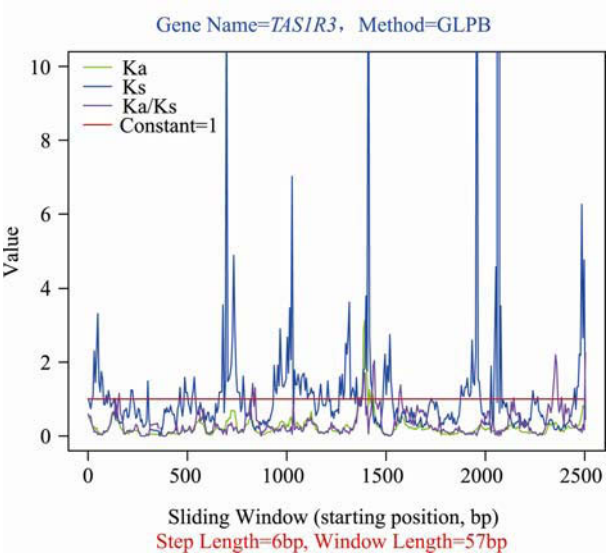

Figure 2 An example for displaying Ka, Ks and Ka/Ks to identify positively selected sites. This analysis was performed based on the *TAS1R3* gene pairs from *Homo sapiens* (NM_152228) and *Canis familiaris* (XM_843615).

similar to the thought of "integral" definition in mathematics. We believe that the software provides an excellent choice when one calculates for positively selected sites. A final note is that we will construct ancestral sequences for the measurement of lineage-specific selective strength in our next update.

Acknowledgements

This work was funded by the National Basic Research Program of China (973 Program) to JY (Grant No. 2006CB910404).

Authors' contributions

DW and JY conceived and designed this study. DW and YZ programmed the software and drafted the manuscript. ZZ supplied several bug reports and modified schemes in the previous version of the software. DW and JZ contributed to data analyses and software testing. JY managed this project 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 Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24: 1586-1591.

2 Tamura, K., *et al.* 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. *Mol. Biol. Evol.* 24: 1596-1599.

3 Librado, P. and Rozas, J. 2009. DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. *Bioinformatics* 25: 1451-1452.

4 Pond, S.L., *et al.* 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21: 676-679.

5 Stajich, J.E., *et al.* 2002. The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.* 12: 1611-1618.

6 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.

7 Zhang, Z. and Yu, J. 2006. Evaluation of six methods for estimating synonymous and nonsynonymous substitution rates. *Genomics Proteomics Bioinformatics* 4: 173-181.

8 Zhang, Z., *et al.* 2006. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. *Genomics Proteomics Bioinformatics* 4: 259-263.

9 Fares, M.A. 2004. SWAPSC: sliding window analysis procedure to detect selective constraints. *Bioinformatics* 20: 2867-2868.

10 Lee, Y.S., *et al.* 2008. WSPMaker: a web tool for calculating selection pressure in proteins and domains using window-sliding. *BMC Bioinformatics* 9: S13.

11 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.

12 Wang, D., *et al.* 2009. How do variable substitution rates influence Ka and Ks calculations? *Genomics Proteomics Bioinformatics* 7: 116-127.

13 Lynn, D.J., *et al.* 2004. Evidence of positively selected sites in mammalian alpha-defensins. *Mol. Biol. Evol.* 21: 819-827.

14 Max, M., *et al.* 2001. Tas1r3, encoding a new candidate taste receptor, is allelic to the sweet responsiveness locus Sac. *Nat. Genet*. 28: 58-63.