User’s Guide

KaKs_Calculator

Version 1.2 (April 2006)

Content

1 Introduction

KaKs_Calculator is a program that calculates nonsynonymous (Ka) and synonymous (Ks) substitution rates through model selection and model averaging. In addition, several currently acknowledged methods for estimating Ka and Ks are also incorporated into it.

The KaKs_Calculator package, including source codes, compiled executables and documentation, is freely available for academic use only at http://evolution.genomics.org.cn/software.htm.

2 Installation

For high efficiency and compatibility with more platforms, the kernel codes of KaKs_Calculator are written in standard C++. For Windows version we use Visual C++ 6.0 for GUI (Graphics User Interface). You can download from the KaKs_Calculator webpage at http://evolution.genomics.org.cn/software.htm the newest package, normally named KaKs_CalculatorXXX.tar.gz (XXX stands for the version).

2.1 Linux/Unix

KaKs_Calculator has been tested on AIX, IRIX and Solaris.

- Unpack the package of KaKs_CalculatorXXX.tar.gz by the following commands.

  ```
  gzip –d KaKs_CalculatorXXX.tar.gz
  tar –xf KaKs_CalculatorXXX.tar
  ```

- If you use other Linux/Unix OS, you have to compile the program in the source codes folder with the help of g++/gcc compiler by yourselves.

  ```
  cd KaKs_CalculatorXXX/src
  make
  ```

2.2 Windows

The Windows version of KaKs_Calculator can run on any IBM compatible computer under Windows Operating System (tested on Windows 2000/XP).

- Unpack the package of KaKs_CalculatorXXX.tar.gz.

- In the folder of “KaKs_CalculatorXXX/bin/Windows/”, just click ‘setup.exe’ for installation.

3 Methods for Calculating Ka and Ks

Calculating Ka and Ks normally involves three steps. Let us assume that the number of lengths between two DNA sequences compared is *n* and the number of substitutions between them is *m*. To calculate Ka and Ks, we need to count the numbers of synonymous (S) and nonsynonymous (N) sites (S + N = *n*) and the numbers of synonymous (S_d) and nonsynonymous (N_d) substitutions ($S_d + N_d = m$). Then it is after correcting multiple substitutions that (N_d/N) and (S_d/S) could represent Ka and Ks, respectively, since the observed number of substitutions underestimates the real number of substitutions as sequences diverge over time. Therefore, we can conclude from mentioned above that these methods normally involve three steps to estimate Ka and Ks: counting S and N, counting S_d and N_d, and correction for multiple substitutions.

Methods for calculating Ka and Ks adopt different substitution models with subtle yet significant differences. They can be classified as approximate methods and maximum-likelihood methods. Different from approximate methods, maximum-likelihood methods adopt the probability theory to finish all three steps mentioned above in one go.

3.1 Approximate Methods

There are several approximate methods incorporated into KaKs_Calculator, and we list their abbreviations in the program and their corresponding reference(s) as follows.

- NG: Nei, M. and Gojobori, T. (1986)
- LWL: Li, W.H., et al. (1985)
- LPB: Li, W.H. (1993) and Pamilo, P. and Bianchi, N.O. (1993)
- MLWL (Modified LWL), MLPB (Modified LPB): Tzeng, Y.H., et al. (2004)
- YN: Yang, Z. and Nielsen, R. (2000)
- MYN (Modified YN): Zhang, Z., et al. (2006)

3.2 Maximum-Likelihood Methods

The method of GY takes account of sequence evolutionary features, such as transition/transversion rate ratio and nucleotide frequencies (reflected in the HKY Model) and incorporates these features into a codon-based model. We extend this method to a set of candidate models in a maximum likelihood framework and use the AIC_c for model selection and model averaging.

- GY: Goldman, N. and Yang, Z. (1994)
- MS (Model Selection), MA (Model Averaging): based on a set of candidate models defined by Posada, D. (2003) as follows.

Model	Substitution Rates	Nucleotide Frequency
JC	$r_{TC}=r_{AG}=r_{TA}=r_{CG}=r_{TG}=r_{CA}$	Equal
F81		Unequal
K2P	$r_{TC}=r_{AG}\neq r_{TA}=r_{CG}=r_{TG}=r_{CA}$	Equal
HKY		Unequal
TrNEF	$r_{TC}\neq r_{AG}\neq r_{TA}=r_{CG}=r_{TG}=r_{CA}$	Equal
TrN		Unequal
K3P	$r_{TC}=r_{AG}\neq r_{TA}=r_{CG}\neq r_{TG}=r_{CA}$	Equal
K3PUF		Unequal
TIMEF	$r_{TC}\neq r_{AG}\neq r_{TA}=r_{CG}\neq r_{TG}=r_{CA}$	Equal
TIM		Unequal
TVMEF	$r_{TC}=r_{AG}\neq r_{TA}\neq r_{CG}\neq r_{TG}\neq r_{CA}$	Equal
TVM		Unequal
SYM	$r_{TC}\neq r_{AG}\neq r_{TA}\neq r_{CG}\neq r_{TG}\neq r_{CA}$	Equal
GTR		Unequal

r_{ij}: substitution rate between i and j, where $i\neq j$ and $i, j\in\{A, C, G, T\}$

4 Format of Sequence

KaKs_Calculator accepts quasi-AXT sequence format as follows. Before calculation, gaps and stop codons between compared sequences will be removed. You can also see "example.axt" in the folder of "KaKs_CalculatorXXX/examples/".

For example:

```
NP_000026
ATGCTCCTGTG-CCACTGGCC
ATCCCC-TGCGCTCACTGGAC

NP_000053
ACAGaTtCTACCc-GCCcACTA--GgtGtt
---ggTTCTCCtACCcA-G-CACTACTggg
```

Each pair of sequences in an axt file contains three lines: a sequence name line and 2 sequence lines. Pairwise sequences are separated from one another by blank lines.

- Sequence name line

```
NP_000026
.
```

- Pairwise sequences lines

```
ATGCTCCTGTG-CCACTGGCC
ATCCCC-TGCGCTCACTGGAC
```

5 Parameters setting

5.1 Linux/Unix

KaKs_Calculator are more suitable for a large number of dataset to calculate Ka and Ks. It reads a pair of sequences and computes corresponding estimates one by one, so that it requires memory proportional to the maximum length among pairwise sequences. In addition, KaKs_Calculator allows user to choose more than one method to calculate Ka and Ks at one running time. The following is the parameters' setting in Linux version.

- -i AXT sequence file name for calculating Ka and Ks
- -o File name for outputting results
- -c Genetic code (Default = 1-Standard Code). For more information about the Genetic Codes, please see the link: http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode=c
- -m Methods for calculating Ka and Ks (Default = MA): NG, LWL, LPB, MLWL, MLPB, YN, MYN, GY, MS, MA, ALL (including all above methods)
- -d File name for details about each candidate model only when using the method of MS or MA
- -h Show help information

For example:

- use MA method and standard code

```
KaKs_Calculator -i test.axt -o test.axt.kaks
```

- use MA method and vertebrate mitochondrial code

```
   KaKs_Calculator -i test.axt -o test.axt.kaks -c 2
```

- use MA method and standard code and output details of model selection on each candidate model

```
   KaKs_Calculator   -i   test.axt   -o   test.axt.kaks   -d
test.axt.details
```

- use LWL, YN and MYN and standard Code

```
   KaKs_Calculator -i test.axt -o test.axt.kaks -m LWL -m YN
-m MYN
```

5.2 Windows

The Windows version provides users with a friendly interface to select input sequences' file, genetic code and method(s) for estimating Ka and Ks. During calculating you can minimize the application window and send it to tray. After finishing calculation, KaKs_Calculator allows users to exports results to file or clipboard at will.

6 Output Format

KaKs_Calculator provides comprehensive information estimated from compared sequences, including numbers of synonymous and nonsynonymous sites, numbers of synonymous and nonsynonymous substitutions, GC contents, maximum-likelihood score, and AIC_C, in addition to synonymous and nonsynonymous substitution rates and their ratio. Meanwhile, Fisher's exact test for small sample is applied to justify the validity of Ka and Ks calculated by these methods.

- Sequence: Name of Pairwise sequence
- Method: Name of method for calculation of Ka and Ks
- Ka: Nonsynonymous substitution rate
- Ks: Synonymous substitution rate
- Ka/Ks: Selective strength
- P-Value(Fisher): The value computed by Fisher exact test
- Length: Sequence length (after removing gaps and stop codon(s))
- S-Sites: Synonymous sites
- N-Sites: Nonsynonymous sites
- Fold-Sites(0:2:4): 0,2,4-fold degenerate sites

- Substitutions: Substitutions between sequences
- S-Substitutions: Synonymous substitutions
- N-Substitutions: Nonsynonymous substitutions
- Fold-S-Substitutions(0:2:4): Synonymous substitutions at 0,2,4-fold
- Fold-N-Substitutions(0:2:4): Nonsynonymous substitutions at 0,2,4-fold
- Divergence-Time: Divergence time
- Substitution-Rate-Ratio(rTC:rAG:rTA:rCG:rTG:rCA/rCA): Ratios of six substitution rates to the substitution rate between C and A
- GC(1:2:3): GC content of entire sequences and of three codon positions
- ML-Score: Maximum likelihood score
- AICc: Value of AICc
- Akaike-Weight: Value of Akaike weight for model selection
- Model: Selected model for the method of MS

7 Acknowledgements

We thank Professor Wen-Hsiung Li for providing us with his computer program and Professor Ziheng Yang for his invaluable source codes in PAML. We are grateful to Heng Li for his advice and Yafeng Hu for his help in software designing. We also thank all anonymous users for reporting bugs and sending suggestions.

8 References

[1] Agresti, A. 1992. A Survey of Exact Inference for Contingency Tables. Statistical Science. 7, 131 -177.

[2] Akaike, H. 1973 Information theory as an extension of the maximum likelihood principle. In Petrov, B.N. and Csaki, F. (eds), Second International Symposium on Information Theory. Akademiai Kiado, Budapest, 267-281

[3] Akaike, H. 1974 A new look at the statistical model identification. IEEE Trans. Autom. Contr. 19, 716-723.

[4] Bierne, N. and Eyre-Walker, A. 2003. The Problem of Counting Sites in the Estimation of the Synonymous and Nonsynonymous Substitution Rates: Implications for the Correlation Between the Synonymous Substitution Rate and Codon Usage Bias. Genetics. 165, 1587-1597.

[5] Burnham, K.P. and Anderson, D.R. 2002 Model Selection and Multimodel Inference: A Practical Information Theoretic Approach. In. Springer-Verlag, New York, 488.

[6] Burnham, K.P. and Anderson, D.R. 2004 Multimodel Inference: Understanding AIC and BIC in Model Selection, Sociological Methods Research, 33, 261-304.

[7] Comeron, J.M. 1999. K-Estimator: calculation of the number of nucleotide substitutions per site and the confidence intervals. Bioinformatics. 15, 763-764.

[8] Gillespie, J.H. 1991. The causes of molecular evolution. Oxford University Press, Oxford, England.

[9] Goldman, N. and Yang, Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. 11, 725-736.

[10] Hasegawa, M., H. Kishino, and T. Yano 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 22, 160-174.

[11] Hurst, L.D. 2002. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends in Genetics. 18, 486-487.

[12] Jukes, T.H., and C. R. Cantor 1969. Evolution of protein molecules, 21-123. In Munro, H.N. eds., Mammalian Protein Metabolism. Academic Press, New York.

[13] Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111-120.

[14] Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, England.

[15] Li, W.H. 1993. Unbiased estimation of the Rates of synonymous and nonsynonymous substitution. J. Mol. Evol. 36, 96-99.

[16] Li, W.H. 1997. Molecular evolution. Sinauer Associates. Sunderland, Mass.

[17] Li, W.H., Wu, C.I. and Luo, C.C. 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.

[18] Muse, S.V. 1996. Estimating synonymous and nonsynonymous substitution rates. Mol. Biol. Evol. 13, 105-114.

[19] Nei, M. and Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3, 418-426.

[20] 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.

[21] Posada, D. 2003 Using Modeltest and PAUP* to select a model of nucleotide substitution. In Baxevanis, A.D. (ed), Current Protocols in Bioinformatics. JohnWiley & Sons, New York, 6.5.1-6.5.14.

[22] 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.

[23] Sullivan, J. and Joyce, P. 2005 Model Selection in Phylogenetics, Annual Review of Ecology, Evolution, and Systematics, 36, 445-466.

[24] Torrents, D., Suyama, M., Zdobnov, E. and Bork, P. 2003. A Genome-Wide Survey of Human Pseudogenes. Genome Res. 13, 2559-2567.

[25] Tzeng, Y.-H., Pan, R. and Li, W.-H. 2004. Comparison of Three Methods for Estimating Rates of Synonymous and Nonsynonymous Nucleotide Substitutions. Mol. Biol. Evol. 21, 2290-2298.

[26] Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. CABIOS. 13, 555-556.

[27] Yang, Z. and Nielsen, R. 2000. Estimating Synonymous and Nonsynonymous Substitution Rates Under Realistic Evolutionary Models. Mol Biol Evol. 17, 32-43.

[28] Zhang, Z., Li, J. and Yu, J. 2006 Computing Ka and Ks with a consideration of unequal transitional substitutions, BMC Evolutionary Biology, 6, 44.

9 Contact

Please send bugs or advice to Zhang Zhang at zhangzhang@genomics.org.cn.