KaKs_Calculator 3.0: Calculating Selective Pressure on Coding and Non-coding Sequences

Zhang Zhang

1 National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
2 CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
3 China National Center for Bioinformation, Beijing 100101, China
4 University of Chinese Academy of Sciences, Beijing 100049, China

Received 24 November 2021; revised 15 December 2021; accepted 21 December 2021
Available online 3 January 2022

Handled by Fangqing Zhao

Abstract KaKs_Calculator 3.0 is an updated toolkit that is capable of calculating selective pressure on both coding and non-coding sequences. Similar to the nonsynonymous/synonymous substitution rate ratio for coding sequences, selection on non-coding sequences can be quantified as the ratio of non-coding nucleotide substitution rate to synonymous substitution rate of adjacent coding sequences. As testified on empirical data, KaKs_Calculator 3.0 shows effectiveness to detect the strength and mode of selection operated on molecular sequences, accordingly demonstrating its great potential to achieve genome-wide scan of natural selection on diverse sequences and identification of potentially functional elements at a whole-genome scale. The package of KaKs_Calculator 3.0 is freely available for academic use only at https://ngdc.cnbc.ac.cn/biocode/tools/BT000001.

Introduction

Detecting natural selection on molecular sequences is of fundamental significance in molecular evolution, comparative genomics, and phylogenetic reconstruction, which can provide profound insights for revealing evolutionary processes of molecular sequences and unveiling complex molecular mechanisms of genome evolution [1]. In principle, estimating selection on DNA sequences requires a reference set of substitutions that is free from selection. As synonymous substitutions do not provoke amino acid changes due to the degeneracy of the genetic code, they are expected to be invisible to selection and thus widely used as a reference that reflects the neutral rate of evolution [2]. Consequently, the ratio of nonsynonymous substitution rate ($K_a$ or $d_{NS}$) to synonymous substitution rate ($K_s$ or $d_S$), namely, $\omega = K_a/K_s$ (or $d_{NS}/d_S$),
is widely adopted to differentiate neutral mutation ($\alpha \approx 1$) from negative (purifying) selection ($\alpha < 1$) and positive (adaptive) selection ($\alpha > 1$), accordingly providing a powerful tool for illuminating molecular evolution of coding sequences (see a popular package in [3]).

Nowadays, a growing body of evidence has shown that non-coding sequences, historically thought as “junk” due to few knowledge on their function relative to coding sequences, are recognized as functional elements to play important regulation roles in multiple biological processes [4] and associate closely with various human diseases [5–7]. Albeit less conserved by comparison with coding sequences, a larger number of non-coding sequences have been identified highly conserved across mammalian genomes [8–10]. Importantly, more non-coding RNA (lncRNA) sequences than previously believed, and particularly, long non-coding RNA (lncRNA) sequences do experience natural selection [11]. As a result, several computational methods have been proposed for the detection of selection acting on non-coding sequences [12], which primarily differ in how to choose a reference of unconstrained evolution, such as, synonymous substitutions of neighboring coding gene [13], intron sequences [14,15], and ancestral repeats [16]. However, there lacks of an implemented algorithm to detect the strength and mode of selective pressure on non-coding sequences, particularly considering an increasing number of non-coding studies conducted worldwide. More importantly, an integrated toolkit that is capable of detecting selection on both coding and non-coding sequences is highly desirable, which would help users achieve genome-wide scan of natural selection on diverse sequences.

Toward this end, here we present KaKs_Calculator 3.0, an updated toolkit for calculating selective pressure on both coding and non-coding sequences. Compared with previous versions [17,18] that focus solely on coding sequences, we implement an algorithm in KaKs_Calculator 3.0 that employs synonymous sites of adjacent coding sequences as a reference to estimate selective pressure acting on non-coding sequences. We test it on empirical data and demonstrate its utility in diagnosing the strength and form of molecular evolution.

Algorithm

The major update of KaKs_Calculator 3.0 is to incorporate an algorithm that is capable of estimating selective pressure on non-coding sequences. Specifically, it uses synonymous substitutions as a reference baseline (similar to [13]), which, albeit thought to be under weak selection [19–21], has been widely adopted for determining the strength and type of selection operated on coding sequences [22–29]. Similar to the Ka/Ks ratio for coding sequences, selective pressure on non-coding sequences ($\xi$) can be quantified as the ratio of non-coding nucleotide substitution rate ($K_n$) to neutral substitution rate (assumed as $K_s$), viz. $\xi = K_n/K_s$, where $K_s$ is inferred from adjacent coding sequences. As the number of observed substitutions is less than the number of real substitutions, we adopt a nucleotide substitution model (e.g., JC/K2P/HKY) to correct multiple substitutions of non-coding sequences. Taking the HKY model [30] as an example, therefore, $K_n$ can be deduced from the observed transitional and transversional substitutions ($S$ and $V$, respectively) as well as four nucleotide frequencies ($\pi_A$, $\pi_T$, $\pi_G$, and $\pi_C$), according to Equation (1) (see Equations 1.27 and 1.28 in [31]).

$$K_n = 2\left(\frac{\pi_T\pi_C}{\pi_V} + \frac{\pi_A\pi_G}{\pi_R}\right) - 2\left(\frac{\pi_T\pi_R}{\pi_V} + \frac{\pi_A\pi_G}{\pi_R}\right) - \frac{\pi_V\pi_R}{\pi_V\pi_R}b$$

where $a = -\log\left(1 - \frac{S}{2\pi_A\pi_G + \pi_R}\right) - \frac{(S\pi_A\pi_G + \pi_R)}{2\pi_A\pi_G + \pi_R}$, $b = -\log\left(1 - \frac{V}{2\pi_A\pi_G + \pi_R}\right)$.

To detect and quantify selection on non-coding sequences, KaKs_Calculator 3.0 provides users with two ways to obtain the value of neutral mutation rate or $K_s$, which is either calculated from adjacent coding sequences uploaded by users or just specified in a straightforward manner by users (Figure 1). As a consequence, KaKs_Calculator 3.0 is capable of detecting selection on both coding and non-coding sequences.

KaKs_Calculator 3.0 is implemented in standard C++ language, enabling higher efficiency and easy compilation on different operation systems (Linux/Windows/Mac). In addition to the new functionality for estimating selection on non-coding sequences as mentioned above, it is also updated by fixing bugs and errors. The package of KaKs_Calculator 3.0, including compiled executables, a Windows application with graphical user interface (GUI), source codes, and example data, accompanying with detailed instructions and documentation, is freely available for academic use only at BioCode (https://ngdc.cnbc.ac.cn/biocode/tools/BT000001), an open-source platform for archiving bioinformatics tools in the National Genomics Data Center (NGDC) [32], China National Center for Bioinformation.

Application on empirical data

To test KaKs_Calculator 3.0, we choose three empirical IncRNA genes that are extensively studied according to LncRNAWiki [7] and collect their human–mouse orthologs as well as their adjacent coding orthologs from NGDC LncBook [33] and National Center of Biotechnology Information (NCBI) RefSeq [34]. Specifically, these non-coding and coding gene symbols with accession numbers are: 1) H19 (NR_002196.2 vs. NR_130973.1) and MRPL23 (NM_021134.4 vs. NM_011288.2); 2) Metastasis-associated lung adenocarcinoma transcript 1 (MALAT1; NR_002819.4 vs. NR_002847.3) and SCYLI (NM_020680.4 vs. NM_01361921.1); and 3) Hox transcript antisense intergenic RNA (HOTAIR; NR_003716.3 vs. NR_047528.1) and HOXC12 (NM_173860.3 vs. NM_00463.2). Based on these orthologous genes, we obtain their corresponding aligned sequences by MAFFT [35] (using parameters: --maxiterate 10000 --localpair).

According to the ratio ($\xi$) of non-coding nucleotide substitution rate to adjacent synonymous substitution rate, we reveal that, although the coding genes undergo strong purifying selection ($\alpha < 1$), these three non-coding genes present diverse selective pressure (Table 1). Strikingly, HOTAIR exhibits positive selection ($\xi > 1$), whereas the rest two genes experience negative selection ($\xi < 1$). HOTAIR is a ~2.3-kb intergenic RNA transcribed from the antisense strand of the HOXC gene cluster [36]. The result of positive selection detected on HOTAIR relative to HOXC12 is consistent well with previous
findings that HOTAIR evolves faster than the neighboring genes [37]. On the contrary, MALAT1, a 8.7-kb non-coding RNA flanked by the highly conserved kinase-like gene SCYL1, is ubiquitously expressed in almost all human tissues, evolutionarily conserved across mammalian species [38], and associated with various cancers [39]. Thus, \( n = 0.464 \) indicates strong selective constraint on MALAT1, in accordance with its physiologic and pathophysiological function [40] and conserved RNA structure [41] as documented by previous studies. Likewise, H19, a 2.3-kb imprinted maternally expressed transcript located near MRPL23, is known for close association with Beckwith-Wiedemann Syndrome and also involved in tumorigenesis [42]. Our result shows that H19 presents stronger selection constraint as indicated by \( \xi = 0.296 \), conforming well with its conserved sequence and structure [43]. It is worth noting that one non-coding sequence may have multiple adjacent coding genes, which are specified by users and thus can lead to different estimates of \( K_s \) and \( \xi \). Taken together, KaKs_Calculator 3.0 is effective in estimating natural selection on non-coding sequences, which has the potential to reveal evolutionarily selective pressures operated on diverse molecular sequences.

In addition, to test the running performance of KaKs_Calculator, we collect an empirical large dataset that contains 15,424 human–mouse orthologous genes retrieved from RefSeq [34] and obtain their codon-based alignments by ParaAT [44] — a parallel tool for constructing multiple protein-coding DNA alignments. KaKs_Calculator 3.0 includes ten computational methods for detecting selection on coding sequences, which fall into approximate methods and maximum-likelihood methods. We choose three approximate methods, NG [23], YN [26], and MYN [27], and one maximum-likelihood method, GY [25], and test on a 64 bit x86 Intel Core i7 machine containing 4 CPU cores with each 3.40 GHz and running Windows 10. For this large-scale data analysis, we find that NG, YN, and MYN all take 2 min and GY takes 11 h, clearly showing that approximate methods are more time-efficient than maximum-likelihood ones. Considering that different users may have different preferences, it should be noted, however, that maximum-likelihood methods are believed to achieve higher accuracy and that different methods adopt different models and strategies and thus can lead to different estimates [45] (see an example in [46] where contradictory findings are produced by different methods).

Table 1 Estimates of selective pressure as well as substitution rates in human–mouse orthologs

| Gene symbol | Kn  | \( \xi = Kn/Ks \) | \( n = Kn/Ks \) | Ka  | Ks  | \( \omega = Ka/Ks \) |
|-------------|-----|------------------|----------------|-----|-----|-------------------|
| H19         | 0.340 | 0.296            | 0.296          | 0.088 | 1.150 | 0.077          |
| MALAT1      | 0.324 | 0.464            | 0.296          | 0.040 | 0.697 | 0.058          |
| HOTAIR      | 0.544 | 1.114            | 0.296          | 0.020 | 0.488 | 0.041          |

Note: \( Ka \), nonsynonymous substitution rate; \( Ks \), synonymous substitution rate; \( Kn \), non-coding nucleotide substitution rate; \( \omega \), selective pressure on coding sequence; \( \xi \), selective pressure on non-coding sequence.
Discussion

KaKs_Calculator 3.0 is significantly updated by achieving the detection of natural selection on non-coding sequences as well as coding sequences. As testified on empirical data, it is of great utility in calculating natural selection on molecular sequences, thus identifying potentially functional elements at a genome-wide scale. Future developments include the detection of selective pressure on small peptides (less than 300 nucleotides) that are encoded by small open reading frames within non-coding sequences [47–49] as well as the implementation of codon-based alignment procedure to help users generate input sequences in an easy-to-use manner.

Code availability

KaKs_Calculator 3.0 is freely available for academic use only at https://ngdc.cncb.ac.cn/biocode/tools/BT000001.

CRediT author statement

Zhang Zhang: Conceptualization, Methodology, Software, Writing - original draft, Writing - review & editing, Funding acquisition, Supervision. The author has read and approved the final manuscript.

Competing interests

The author has declared no competing interests.

Acknowledgments

I would like to extend special thanks to Lina Ma for constructive suggestions and discussions on this work and Zhao Li for valuable help on data collection and test. I also thank Zhuming Fan for designing the logo as well as Qing Guo and Lin Dai for fixing a bug on Windows GUI. I am extremely grateful to a number of users for reporting bugs and sending comments since the first release of KaKs_Calculator in 2006. This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (Grant No. XDA19050302), the National Natural Science Foundation of China (Grant Nos. 31871328 and 32030021), the National Key R&D Program of China (Grant No. 2017YFC0907502), and the International Partnership Program of the Chinese Academy of Sciences (Grant No. 153F11KYSB20160008).

ORCID

ORCID 0000-0001-6603-5060 (Zhang Zhang)

References

[1] Li WH. Molecular evolution. Sunderland (Massachusetts): Sinauer Associates; 1997.
[2] Hurst LD. The Ka/Ks ratio: diagnosing the form of sequence evolution. Trends Genet 2002;18:486–7.
[3] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 2007;24:1586–91.
[4] Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
[5] Luo H, Bu D, Shao L, Li Y, Sun L, Wang C, et al. Single-cell long non-coding RNA landscape of T cells in human cancer immunity. Genomics Proteomics Bioinformatics 2021;19:377–93.
[6] Anastasiadou E, Jacob LS, Slack FJ. Non-coding RNA networks in cancer. Nat Rev Cancer 2018;18:5–18.
[7] Liu L, Li Z, Liu C, Zou D, Li Q, Feng C, et al. LncRNAWiki 2.0: a knowledgebase of human long non-coding RNAs with enhanced curated model and database system. Nucleic Acids Res 2022;50:D190–5.
[8] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, et al. Ultraconserved elements in the human genome. Science 2004;304:1321–5.
[9] Habic A, Mattick JS, Calin GA, Krese R, Konc J, Kunej T. Genetic variations of ultraconserved elements in the human genome. OMICS 2019;23:549–59.
[10] Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 2009;458:223–7.
[11] Diederichs S. The four dimensions of noncoding RNA conservation. Trends Genet 2014;30:121–3.
[12] Zhen Y, Andolfatto P. Methods to detect selection on noncoding DNA. Methods Mol Biol 2012;856:141–59.
[13] Wang WSW, Nielsen R. Detecting selection in noncoding regions of nucleotide sequences. Genetics 2004;167:949–58.
[14] Parsch J, Novozhilov S, Samadin-Peter SS, Wong KM, Andolfatto P. On the utility of short intron sequences as a reference for the detection of positive and negative selection in Drosophila. Mol Biol Evol 2010;27:1226–34.
[15] Hoffman MM, Birney E. Estimating the neutral rate of nucleotide substitution using introns. Mol Biol Evol 2007;24:522–31.
[16] Bush EC, Lahn BT. A genome-wide screen for noncoding elements important in primate evolution. BMC Evol Biol 2008;8:17.
[17] Zhang Z, Li J, Zhao XQ, Wang J, Wong GS, Yu J. KaKs_Calculator: calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinformatics 2006;4:259–63.
[18] Wang D, Zhang Y, Zhang Z, Zhu J, Yu J. KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics 2010;8:77–80.
[19] Shabalina SA, Spiridonov NA, Kashina A. Sounds of silence: synonymous nucleotides as a key to biological regulation and complexity. Nucleic Acids Res 2013;41:2073–94.
[20] Hershberg R, Petrov DA. Selection on codon bias. Annu Rev Genet 2008;42:287–99.
[21] Plotkin JB, Kudla G. Synonymous but not the same: the causes and consequences of codon bias. Nat Rev Genet 2011;12:32–42.
[22] Li WH, Wu CI, Luo CC. A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 1985;2:150–74.
[23] Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 1986;3:418–26.
[24] Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J Mol Evol 1993;36:96–9.
[25] Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 1994;11:725–36.
[26] Yang Z, Nielsen R. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. Mol Biol Evol 2000;17:32–43.
