BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 5 2013, pages 645–646
doi:10.1093/bioinformatics/btt009

Genome analysis

Advance Access publication January 12, 2013

gKaKs: the pipeline for genome-level Ka/Ks calculation

Chengjun Zhang[1,†], Jun Wang[2,†], Manyuan Long[1] and Chuanzhu Fan[2,*]

[1]Department of Ecology and Evolution, University of Chicago, Chicago, IL 60637 and [2]Department of Biological Sciences, Wayne State University, Detroit, MI 48202, USA

Associate Editor: Martin Bishop

ABSTRACT

Summary: gKaKs is a codon-based genome-level Ka/Ks computation pipeline developed and based on programs from four widely used packages: BLAT, BLASTALL (including bl2seq, formatdb and fastacmd), PAML (including codeml and yn00) and KaKs_Calculator (including 10 substitution rate estimation methods). gKaKs can automatically detect and eliminate frameshift mutations and premature stop codons to compute the substitution rates (Ka, Ks and Ka/Ks) between a well-annotated genome and a non-annotated genome or even a poorly assembled scaffold dataset. It is especially useful for newly sequenced genomes that have not been well annotated. We applied gKaKs to estimate the genome-wide substitution rates in five pairs of closely related species. The average Ka and Ks computed by gKaKs were consistent with previous studies. We also compared the Ka, Ks and Ka/Ks of mouse and rat orthologous protein-coding genes estimated by gKaKs and based on the alignments generated by PAL2NAL. Results from two methods are compatible.

Availability and implementation: gKaKs is implemented in Perl and is freely available on http://longlab.uchicago.edu/?q=gKaKs. The detailed user manual is available on the website.

Contact: cfan@wayne.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 19, 2012; revised on January 3, 2013; accepted on January 4, 2013

1 INTRODUCTION

Open reading frames (ORFs) are embedded in genomic DNA sequences. Once detected, ORF sequences can be aligned at either DNA or protein levels. Because of the degeneracy of genetic codes, protein sequences are more conserved than DNA sequences. Therefore, protein sequences can be used to compare two ORFs with greater evolutionary distance and provide fundamental biological and evolutionary information of DNA sequences.

Previously, sequence alignment programs for ORFs, such as Bioperl toolkit (Stajich *et al.*, 2002), RevTrans (Wernersson and Pedersen, 2003), transAlign (Bininda-Emonds, 2005), PAL2NAL (Suyama *et al.*, 2006), TranslatorX (Abascal *et al.*, 2010) and MACSE (Ranwez *et al.*, 2011), use a consecutive three-step approach: (i) translation of nucleotide sequences into protein sequences; (ii) alignment of protein sequences; and (iii) alignment of DNA sequences according to protein alignment. However, this three-step alignment approach is highly sensitive to frameshift and non-sense mutations. Such mutations are common in genomic datasets, and they can be the result of either sequencing errors or true evolutionary events leading to loss of function or neo-functionalized gene copies (Ranwez *et al.*, 2011). In addition, the three-step alignment approach also creates a problem when the aligned sequences do not have annotated ORFs, or homologous genes have only partial sequences originating from a common ancestor. In this study, we present an alternative 'codon-based alignment' approach where DNA sequences are directly aligned while also considering codon information. Codon-aware alignment based on BLAT and bl2seq programs automatically aligns coding sequences (CDSs) from a reference genome to another genome sequences, deleting the non-homologous sequences such as those existing in a pair of partial duplicated genes. It automatically handles the existence of multiple frameshift mutations and/or early stop codons in both input DNA sequences. After the alignment, the pipeline uses codeml/yn00 in PAML (Yang, 2007) or other 10 algorithms in KaKs_Calculator (Zhang *et al.*, 2006) to compute the Ka, Ks and Ka/Ks of the aligned homologous sequence pairs, which can be based on >20 genetic codon models. We named the whole pipeline 'gKaKs'. In terminology of gKaKs pipeline, Ka, Ks and Ka/Ks are equivalent to dN, dS and dN/dS, respectively.

2 IMPLEMENTATION

gKaKs uses programs from four widely used bioinformatics packages: BLAT (Kent, 2002), bl2seq, formatdb and fastacmd from BLASTALL (Altschul *et al.*, 1990), codeml/yn00 from PAML (Yang, 2007) and KaKs_Calculator (Zhang *et al.*, 2006). gKaKs inputs data from two genomes, both data in a 'fasta' file format. The reference genome sequences must include CDSs and have an accompanying '.gff/gtf' file that contains the CDS coordinate information. The target genome sequences can be either CDSs or genomic DNA sequences. Implementation of gKaKs includes four basic steps: (i) identification of homologues to reference species genes via BLAT; (ii) alignment of each exon CDS of the reference species' genes to the homologous sequences in the target genome via bl2seq; (iii) consolidation of all the exon CDS alignments together, deletion of non-triplet indels/early stop codons and recording of the alignments that have frameshift mutations, stop codons and other errors; and (iv) calculation of the Ka, Ks and Ka/Ks for each homologous sequence pair via codeml/yn00 or other 10 methods in KaKs_Calculator. In addition to genome-wide analysis, gKaKs can also be used to

*To whom correspondence should be addressed.
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

compute Ka, Ks and Ka/Ks values for a list of known homologous gene pairs. The implementation of gKaKs for a list of homologous gene pairs can be carried out using the file with the gene names of homologous genes pairs, two 'fasta' files containing the reference CDSs and target CDSs and the '.gtf/gff' file that records coordinates of the reference CDSs. Further, one caution about gKaKs is that the process of BLAT can generate a large number of gaps if two genomes are highly divergent; therefore, gKaKs pipeline performs better for comparison of closely related species with sequence similarity at least 70%. More detail about gKaKs pipeline is shown in Supplementary Figure S1 and at the website http://longlab.uchicago.edu/?q=gKaKs.

We tested gKaKs in five species pairs: human versus chimpanzee, mouse versus rat, *Drosophila melanogaster* versus *Drosophila simulans*, *Oryza sativa* versus *Oryza glaberrima*, and *Arabidopsis thaliana* versus *Arabidopsis lyrata* (Supplementary Table S1). For each pair, the former species was treated as the reference genome of which both the CDSs and the '.gff/gtf' file were used. The latter species was treated as the non-annotated genome of which only the genome sequences were used. We used genome sequences from five chromosomes and implemented gKaKs pipeline for substitution rate estimations. We estimated Ka and Ks using four methods: codeml (Goldman and Yang, 1994), yn00 (Yang, 2007), NG (Nei and Gojobori, 1986) and LWL (Li *et al.*, 1985). The distribution of Ka, Ks and Ka/Ks values are shown in Supplementary Figure S2. The average Ks values of these five pair comparisons based on the codeml method are 0.0292, 0.1747, 0.1189, 0.0572 and 0.1394, respectively. The average Ka values of these five pair comparisons based on the codeml method are 0.0097, 0.0250, 0.0146, 0.0180 and 0.0294, respectively. These average Ka and Ks values are consistent with previous studies (Supplementary Table S2; Andrea *et al.*, 2002; Hughes and Friedman, 2009; Wolf *et al.*, 2009; Yang and Gaut, 2011). Based on Ka/Ks, we further identified genes that potentially evolved under positive selection (with $Ka/Ks > 1$). Overall, we identified 1400 gene pairs in human–chimpanzee, 172 gene pairs in mouse–rat, 137 gene pairs in *D.melanogaster–D.simulans*, 4463 gene pairs in *O.sativa–O.glaberrima* and 478 gene pairs in *A.thaliana–A.lyrata*, which might be driven by positive selection.

To further demonstrate the reliability of gKaKs pipeline, we estimated the substitution rates of mouse and rat orthologous genes using alignment from gKaKs and PAL2NAL. PAL2NAL is one of most popular programs for handling multiple ORF sequence comparisons and generating output files for downstream Ka/Ks calculation. However, PAL2NAL is not convenient tool to handle large-scale whole-genome sequence data, and it cannot efficiently align pseudogenes with the presences of frameshift and non-sense mutations. We downloaded the mouse and rat orthologous genes that were annotated by Ensembl Biomart. We only selected ~14 700 protein-coding 'one orthologue to one orthologue' genes. codeml was used to estimate the substitution rates for both alignments. As shown in Supplementary Table S3, estimations between gKaKs and PAL2NAL are compatible, whereas values from gKaKs are smaller than those from PAL2NAL. The difference largely resulted from gKaKs alignment, removing the non-homologous sequences of the orthologous genes whose substitution rates are relatively high.

3 CONCLUSION

With the advancements of next-generation sequencing technology, large numbers of genome sequences are becoming available rapidly. Comparative genomics analysis between closely related model species with well-annotated genomes and non-model species will provide enormous benefit for functional annotation of non-model species genomes. gKaKs is a useful tool to compute nucleotide substitution rates between genes in well-annotated genomes and their homologous sequences in non-annotated genomes. The results can (i) be used to compute the level of sequence divergence between two species through estimating average Ks and substitution rate; (ii) be used to estimate the number of orthologous/paralogous gene pairs under functional constraints or driven by the positive selection; and (iii) be used as evidence for gene annotation.

ACKNOWLEDGEMENT

The authors thank Dr Edward M. Golenberg and three anonymous reviewers for valuable comments and suggestions for pipeline optimization and manuscript improvement.

Funding: This work was supported by start-up fund from Wayne State University to C.F. and National Science Foundation grant (MCB1026200) to M.L.

Conflict of Interest: none declared.

REFERENCES

Abascal,F. *et al.* (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. *Nucleic Acids Res.*, **38**, W7–W13.

Altschul,S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Andrea,J.B. *et al.* (2002) A test for faster X evolution in *Drosophila*. *Mol. Biol. Evol.*, **19**, 1816–1819.

Bininda-Emonds,O.R.P. (2005) transAlign: using amino acids to facilitate the multiple alignment of protein-coding DNA sequences. *BMC Bioinformatics*, **6**, 156.

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

Kent,W.J. (2002) BLAT—the BLAST-like alignment tool. *Genome Res.*, **12**, 656–664.

Hughes,A.L. and Friedman,R. (2009) More radical amino acid replacements in primates than in rodents: support for the evolutionary role of effective population size. *Gene*, **440**, 50–56.

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.

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

Ranwez,V. *et al.* (2011) MACSE: multiple alignment of coding SEquences accounting for frameshifts and stop codons. *PloS One*, **6**, e22594.

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

Suyama,M. *et al.* (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.*, **34**, W609–W612.

Wernersson,R. and Pedersen,A.G. (2003) RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. *Nucleic Acids Res.*, **31**, 3537–3539.

Wolf,J.B.W. *et al.* (2009) Nonlinear dynamics of nonsynonymous (d_N) and synonymous (d_S) substitution rates affects inference of selection. *Genome Biol. Evol.*, **1**, 308–319.

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

Yang,L. and Gaut,B. (2011) Factors that contribute to variation in evolutionary rate among *Arabidopsis* genes. *Mol. Biol. Evol.*, **28**, 2359–2369.

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