WIKIPEDIA

K_a/K_s ratio

In genetics, the **K_a/K_s ratio**, also known as **ω** or ***d_N/d_S* ratio**,[a] is used to estimate the balance between neutral mutations, purifying selection and beneficial mutations acting on a set of homologous protein-coding genes. It is calculated as the ratio of the number of nonsynonymous substitutions per non-synonymous site (K_a), in a given period of time, to the number of synonymous substitutions per synonymous site (K_s), in the same period. The latter are assumed to be neutral, so that the ratio indicates the net balance between deleterious and beneficial mutations. Values of K_a/K_s significantly above 1 are unlikely to occur without at least some of the mutations being advantageous. If beneficial mutations are assumed to make little contribution, then K_s estimates the degree of evolutionary constraint.

Contents

Context

Selection acts on variation in phenotypes, which are often the result of mutations in protein-coding genes. The genetic code is written in DNA sequences as codons, groups of three nucleotides. Each codon represents a single amino acid in a protein chain. However, there are more codons (64) than amino acids found in proteins (20), so many codons are effectively synonyms. For example, the DNA codons TTT and TTC both code for the amino acid Phenylalanine, so a change from the third T to C makes no difference to the resulting protein. On the other hand, the codon GAG codes for Glutamic acid while the codon GTG codes for Valine, so a change from the middle A to T does change the resulting protein, for better or (more likely) worse,[b] so the change is not a synonym. These changes are illustrated in the tables below.

The K_a/K_s ratio measures the relative rates of synonymous and nonsynonymous substitutions at a particular site.

A point mutation causing a synonymous substitution

Type of structure	Before	Change	After	Result
Codon in a DNA sequence	TTT	harmless mutation;[c] Synonymous substitution	TTC	
↓ codes for	↓ codes for		↓ codes for	
Amino acid in a Protein	Phenylalanine	no change	Phenylalanine	Normal protein, normal function

A point mutation causing a nonsynonymous substitution

Type of structure	Before	Change	After	Result
Codon in a DNA sequence	GAG	Missense mutation; Nonsynonymous substitution	GTG	
↓ codes for	↓ codes for		↓ codes for	
Amino acid in a Protein	Glutamic acid	structural change	Valine	Altered protein may or may not cause harm (e.g. disease) or give new advantage

Methods

Methods for estimating K_a and K_s use a sequence alignment of two or more nucleotide sequences of homologous genes that code for proteins (rather than being genetic switches, controlling development or the rate of activity of other genes). Methods can be classified into three groups: approximate methods, maximum-likelihood methods, and counting methods. However, unless the sequences to be compared are distantly related (in which case maximum-likelihood methods prevail), the class of method used makes a minimal impact on the results obtained; more important are the assumptions implicit in the chosen method.[1]:498

Approximate methods

Approximate methods involve three basic steps: (1) counting the number of synonymous and nonsynonymous sites in the two sequences, or estimating this number by multiplying the sequence length by the proportion of each class of substitution; (2) counting the number of synonymous and nonsynonymous substitutions; and (3) correcting for multiple substitutions.

These steps, particularly the latter, require simplistic assumptions to be made if they are to be achieved computationally; for reasons discussed later, it is impossible to exactly determine the number of multiple substitutions.[1]

Maximum-likelihood methods

The maximum-likelihood approach uses probability theory to complete all three steps simultaneously.[1] It estimates critical parameters, including the divergence between sequences and the transition/transversion ratio, by deducing the most likely values to produce the input data.[1]

Counting methods

In order to quantify the number of substitutions, one may reconstruct the ancestral sequence and record the inferred changes at sites (straight counting – likely to provide an underestimate); fitting the substitution rates at sites into predetermined categories (Bayesian approach; poor for small data sets); and generating an individual substitution rate for each codon (computationally expensive). Given enough data, all three of these approaches will tend to the same result.[2]

Interpreting results

The K_a/K_s ratio is used to infer the direction and magnitude of natural selection acting on protein coding genes. A ratio greater than 1 implies positive or Darwinian selection (driving change); less than 1 implies purifying or stabilizing selection (acting against change); and a ratio of exactly 1 indicates neutral (i.e. no) selection. However, a combination of positive and purifying selection at different points within the gene or at different times along its evolution may cancel each other out. The resulting averaged value can mask the presence of one of the selections and lower the seeming magnitude of another selection.

Of course, it is necessary to perform a statistical analysis to determine whether a result is significantly different from 1, or whether any apparent difference may occur as a result of a limited data set. The appropriate statistical test for an approximate method involves approximating dN − dS with a normal approximation, and determining whether 0 falls within the central region of the approximation. More sophisticated likelihood techniques can be used to analyse the results of a Maximum Likelihood analysis, by performing a chi-squared test to distinguish between a null model ($K_a/K_s = 1$) and the observed results.[1]

Utility

The K_a/K_s ratio is a more powerful test of the neutral model of evolution than many others available in population genetics as it requires fewer assumptions.[1]

Complications

There is often a systematic bias in the frequency at which various nucleotides are swapped, as certain mutations are more probable than others.[1] For instance, some lineages may swap C to T more frequently than they swap C to A. In the case of the amino acid Asparagine, which is coded by the codons AAT or AAC, a high C->T exchange rate will increase the proportion of synonymous substitutions at this codon, whereas a high C→A exchange rate will increase the rate of non-synonymous substitutions. Because it is rather common for transitions (T⇔C & A⇔G) to be favoured over transversions (other changes),[1] models must account for the possibility of non-homogeneous rates of exchange.[3] Some simpler approximate methods, such as those of Miyata & Yasunaga and Nei & Gojobori, neglect to take these into account, which generates a faster computational time at the expense of accuracy; these methods will systematically overestimate N and underestimate S.[1]

Further, there may be a bias in which certain codons are preferred in a gene, as a certain combination of codons may improve translational efficiency.[1]

In addition, as time progresses, it is possible for a site to undergo multiple modifications. For instance, a codon may switch from AAA→AAC→AAT→AAA. There is no way of detecting multiple substitutions at a single site, thus the estimate of the number of substitutions is always an underestimate. In addition, in the example above two non-synonymous and one synonymous substitution occurred at the third site; however, because substitutions restored the original sequence, there is no evidence of any substitution. As the divergence time between two sequences increases, so too does the amount of multiple substitutions. Thus "long branches" in a dN/dS analysis can lead to underestimates of both dN and dS, and the longer the branch, the harder it is to correct for the introduced noise.[3] Of course, the ancestral sequence is usually unknown, and two lineages being compared will have been evolving in parallel since their last common ancestor. This effect can be mitigated by constructing the ancestral sequence; the accuracy of this sequence is enhanced by having a large number of sequences descended from that common ancestor to constrain its sequence by phylogenetic methods.[1]

Methods that account for biases in codon usage and transition/transversion rates are substantially more reliable than those that do not.[1]

Limitations

Although the K_a/K_s ratio is a good indicator of selective pressure at the sequence level, evolutionary change can often take place in the regulatory region of a gene which affects the level, timing or location of gene expression. K_a/K_s analysis will not detect such change. It will only calculate selective pressure within protein coding regions. In addition, selection that does not cause differences at an amino acid level—for instance, balancing selection—cannot be detected by these techniques.[1]

Another issue is that heterogeneity within a gene can make a result hard to interpret. For example, if $K_a/K_s = 1$, it could be due to relaxed selection, or to a chimera of positive and purifying selection at the locus. A solution to this limitation would be to apply K_a/K_s analysis across many species at individual codons.

The K_a/K_s method requires a rather strong signal in order to detect selection. In order to detect selection between lineages, then the selection, averaged over all sites in the sequence, must produce a K_a/K_s greater than one—quite a feat if regions of the gene are strongly conserved. In order to detect selection at specific sites, then the K_a/K_s ratio must be greater than one when averaged over all included *lineages* at that site—implying that the site must be under selective pressure in all sampled lineages. This limitation can be moderated by allowing the K_a/K_s rate to take multiple values across sites and across lineages; the inclusion of more lineages also increases the power of a sites-based approach.[1]

Further, the method lacks the capability to distinguish between positive and negative nonsynonymous substitutions. Some amino acids are chemically similar to one another, whereas other substitutions may place an amino acid with wildly different properties to its precursor. In most situations, a smaller chemical change is more likely to allow the protein to continue to function, and a large chemical change is likely to disrupt the chemical structure and cause the protein to malfunction. However, incorporating this into a model is not straightforward as the relationship between a nucleotide substitution and the effects of the modified chemical properties is very difficult to determine.[1]

An additional concern is that the effects of time must be incorporated into an analysis, if the lineages being compared are closely related; this is because it can take a number of generations for natural selection to "weed out" deleterious mutations from a population, especially if their effect on fitness is weak.[4][5][6][7] This limits the usefulness of the K_a/K_s ratio for comparing closely related populations.

Individual codon approach

Additional information can be gleaned by determining the K_a/K_s ratio at specific codons within a gene sequence. For instance, the frequency-tuning region of an opsin may be under enhanced selective pressure when a species colonises and adapts to new environment, whereas the region responsible for initializing a nerve signal may be under purifying selection. In order to detect such effects, one would ideally calculate the K_a/K_s ratio at each site. However this is computationally expensive and in practise, a number of K_a/K_s classes are established, and each site is shoehorned into the best-fitting class.[1]

The first step in identifying whether positive selection acts on sites is to compare a test where the K_a/K_s ratio is constrained to be < 1 in all sites to one where it may take any value, and see if permitting K_a/K_s to exceed 1 in some sites improves the fit of the model. If this is the case, then sites fitting into the class where $K_a/K_s > 1$ are candidates to be experiencing positive selection. This form of test can either identify sites that further laboratory research can examine to determine possible selective pressure; or, sites believed to have functional significance can be assigned into different K_a/K_s classes before the model is run.[1]

Notes

a. The terms K_a/K_S and d_N/d_S are used interchangeably. Note however that D_n and D_s are different parameters from d_N and d_S (or K_A and K_S). D_n and D_s are count estimates, which represent the total numbers of non-synonymous and synonymous substitutions.
b. "Better" means that the change is advantageous and will be selected for by natural selection. "Worse" means that the change is harmful, and will be selected against.
c. Often but not always a "silent mutation".

References

1. Yang, Z.; Bielawski, J. P. (2000). "Statistical methods for detecting molecular adaptation". *Trends in Ecology & Evolution*. **15** (12): 496–503. CiteSeerX 10.1.1.19.6537 (https://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.19.6537). doi:10.1016/S0169-5347(00)01994-7 (https://doi.org/10.1016%2FS0169-5347%2800%2901994-7). PMID 11114436 (https://pubmed.ncbi.nlm.nih.gov/11114436).
2. Kosakovsky Pond, S. L.; Frost, S. D. W. (2005). "Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection" (https://doi.org/10.1093/molbev/msi105). *Molecular Biology and Evolution*. **22** (5): 1208–22. doi:10.1093/molbev/msi105 (https://doi.org/10.1093%2Fmolbev%2Fmsi105). PMID 15703242 (https://pubmed.ncbi.nlm.nih.gov/15703242).
3. Hurst, L. (2002). "The Ka/Ks ratio: diagnosing the form of sequence evolution". *Trends in Genetics*. **18** (9): 486–489. doi:10.1016/S0168-9525(02)02722-1 (https://doi.org/10.1016%2FS0168-9525%2802%2902722-1). PMID 12175810 (https://pubmed.ncbi.nlm.nih.gov/12175810).
4. Rocha, E. P. C.; Smith, J. M.; Hurst, L. D.; Holden, M. T. G.; Cooper, J. E.; Smith, N. H.; Feil, E. J. (2006). "Comparisons of dN/dS are time dependent for closely related bacterial genomes". *Journal of Theoretical Biology*. **239** (2): 226–35. doi:10.1016/j.jtbi.2005.08.037 (https://doi.org/10.1016%2Fj.jtbi.2005.08.037). PMID 16239014 (https://pubmed.ncbi.nlm.nih.gov/16239014).
5. Kryazhimskiy S, Plotkin JB (2008). "The Population Genetics of dN/dS" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2596312). *PLoS Genetics*. **4** (12): e1000304. doi:10.1371/journal.pgen.1000304 (https://doi.org/10.1371%2Fjournal.pgen.1000304). PMC 2596312 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2596312). PMID 19081788 (https://pubmed.ncbi.nlm.nih.gov/19081788).
6. Peterson GI, Masel J (2009). "Quantitative Prediction of Molecular Clock and Ka/Ks at Short Timescales" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2912

466). *Molecular Biology & Evolution*. **26** (11): 2595–2603. doi:10.1093/molbev/msp175 (https://doi.org/10.1093%2Fmolbev%2Fmsp175). PMC 2912466 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2912466). PMID 19661199 (https://pubmed.ncbi.nlm.nih.gov/19661199).

7. Mugal, CF; Wolf JBW; Kaj I (2014). "Why Time Matters: Codon Evolution and the Temporal Dynamics of dN/dS" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3879453). *Molecular Biology and Evolution*. **31** (1): 212–231. doi:10.1093/molbev/mst192 (https://doi.org/10.1093%2Fmolbev%2Fmst192). PMC 3879453 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3879453). PMID 24129904 (https://pubmed.ncbi.nlm.nih.gov/24129904).

Further reading

- Li, WH; Wu, CI; Luo, CC (March 1985). "A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes" (https://doi.org/10.1093/oxfordjournals.molbev.a040343). *Mol. Biol. Evol*. **2** (2): 150–74. doi:10.1093/oxfordjournals.molbev.a040343 (https://doi.org/10.1093%2Foxfordjournals.molbev.a040343). PMID 3916709 (https://pubmed.ncbi.nlm.nih.gov/3916709).
- Nei M, Gojobori T (September 1986). "Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions" (https://doi.org/10.1093/oxfordjournals.molbev.a040410). *Mol. Biol. Evol*. **3** (5): 418–26. doi:10.1093/oxfordjournals.molbev.a040410 (https://doi.org/10.1093%2Foxfordjournals.molbev.a040410). PMID 3444411 (https://pubmed.ncbi.nlm.nih.gov/3444411).
- Li WH (January 1993). "Unbiased estimation of the rates of synonymous and nonsynonymous substitution". *J. Mol. Evol*. **36** (1): 96–9. doi:10.1007/bf02407308 (https://doi.org/10.1007%2Fbf02407308). PMID 8433381 (https://pubmed.ncbi.nlm.nih.gov/8433381).
- Pamilo P, Bianchi NO (March 1993). "Evolution of the Zfx and Zfy genes: rates and interdependence between the genes" (https://doi.org/10.1093/oxfordjournals.molbev.a040003). *Mol. Biol. Evol*. **10** (2): 271–81. doi:10.1093/oxfordjournals.molbev.a040003 (https://doi.org/10.1093%2Foxfordjournals.molbev.a040003). PMID 8487630 (https://pubmed.ncbi.nlm.nih.gov/8487630).
- Muse SV, Gaut BS (September 1994). "A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome" (https://doi.org/10.1093/oxfordjournals.molbev.a040152). *Mol. Biol. Evol*. **11** (5): 715–24. doi:10.1093/oxfordjournals.molbev.a040152 (https://doi.org/10.1093%2Foxfordjournals.molbev.a040152). PMID 7968485 (https://pubmed.ncbi.nlm.nih.gov/7968485).
- Goldman N, Yang Z (September 1994). "A codon-based model of nucleotide substitution for protein-coding DNA sequences" (https://doi.org/10.1093/oxfordj

ournals.molbev.a040153). *Mol. Biol. Evol*. **11** (5): 725–36. doi:10.1093/oxfordjournals.molbev.a040153 (https://doi.org/10.1093%2Foxfordjournals.molbev.a040153). PMID 7968486 (https://pubmed.ncbi.nlm.nih.gov/7968486).

- Comeron JM (December 1995). "A method for estimating the numbers of synonymous and nonsynonymous substitutions per site". *J. Mol. Evol*. **41** (6): 1152–9. doi:10.1007/bf00173196 (https://doi.org/10.1007%2Fbf00173196). PMID 8587111 (https://pubmed.ncbi.nlm.nih.gov/8587111).
- Ina Y (February 1995). "New methods for estimating the numbers of synonymous and nonsynonymous substitutions". *J. Mol. Evol*. **40** (2): 190–226. doi:10.1007/bf00167113 (https://doi.org/10.1007%2Fbf00167113). PMID 7699723 (https://pubmed.ncbi.nlm.nih.gov/7699723).
- Yang Z (October 1997). "PAML: a program package for phylogenetic analysis by maximum likelihood" (https://doi.org/10.1093/bioinformatics/13.5.555). *Comput. Appl. Biosci*. **13** (5): 555–6. doi:10.1093/bioinformatics/13.5.555 (https://doi.org/10.1093%2Fbioinformatics%2F13.5.555). PMID 9367129 (https://pubmed.ncbi.nlm.nih.gov/9367129).
- Yang Z, Nielsen R (January 2000). "Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models" (https://doi.org/10.1093/oxfordjournals.molbev.a026236). *Mol. Biol. Evol*. **17** (1): 32–43. doi:10.1093/oxfordjournals.molbev.a026236 (https://doi.org/10.1093%2Foxfordjournals.molbev.a026236). PMID 10666704 (https://pubmed.ncbi.nlm.nih.gov/10666704).
- Zhang Z, Li J, Yu J (2006). "Computing Ka and Ks with a consideration of unequal transitional substitutions" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1552089). *BMC Evol. Biol*. **6** (1): 44. doi:10.1186/1471-2148-6-44 (https://doi.org/10.1186%2F1471-2148-6-44). PMC 1552089 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1552089). PMID 16740169 (https://pubmed.ncbi.nlm.nih.gov/16740169).
- Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, Yu J (November 2006). "KaKs_Calculator: calculating Ka and Ks through model selection and model averaging" (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5054075). *Genomics Proteomics Bioinformatics*. **4** (4): 259–63. doi:10.1016/S1672-0229(07)60007-2 (https://doi.org/10.1016%2FS1672-0229%2807%2960007-2). PMC 5054075 (https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5054075). PMID 17531802 (https://pubmed.ncbi.nlm.nih.gov/17531802).

- For a simple introduction, see Hurst, L. (2002). "The Ka/Ks ratio: diagnosing the form of sequence evolution". *Trends in Genetics*. **18** (9): 486–489. doi:10.1016/S0168-9525(02)02722-1 (https://doi.org/10.1016%2FS0168-9525%2802%2902722-1). PMID 12175810 (https://pubmed.ncbi.nlm.nih.gov/12175810).

External links

- KaKs_Calculator (https://code.google.com/p/kaks-calculator)
- Free online server tool that calculates KaKs ratios among multiple sequences (http://services.cbu.uib.no/tools/kaks)
- SeqinR: A free and open biological sequence analysis package for the R language that includes KaKs calculation (https://cran.r-project.org/web/packages/seqinr/index.html)

Retrieved from "https://en.wikipedia.org/w/index.php?title=Ka/Ks_ratio&oldid=965144522"

This page was last edited on 29 June 2020, at 17:29 (UTC).