J Mol Evol (2003) 56:689–690
DOI: 10.1007/s00239-002-2440-z

JOURNAL OF MOLECULAR EVOLUTION

Appendix to Paper by Wall and Herbeck

Montgomery Slatkin, John Novembre

Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720-3140, USA

Received: 29 January 2002 / Accepted: 16 December 2002

Abstract. A method is described for determining the number of preferred codons in taxa in which G + C levels differ. If the hypothesis of random codon usage is not rejected, there are no preferred codons. If that hypothesis is rejected, then a model with one or two preferred codons is fitted to the data and a likelihood ratio test is used to determine whether there are one or two preferred codons. A C++ program is freely available to perform the calculations.

Key words: Codon usage bias — G + C content

The study by Wall and Herbeck is based on comparing the extent of codon usage bias in a wide range of taxa. Codon usage may reflect the G + C content in the genome or may differ significantly from that expected from the genomic G + C content. To consider the evolution of codon usage bias, it is necessary to quantify the extent of bias not attributable to G + C content. To do so, the following procedure was used. For a codon with k-fold redundancy ($k = 2, 3, 4$, or 6), e_k is the expected codon usage if usage was determined by f, the proportion of G + C in the genome. The e_i depend on f and the genetic code. In all cases, the expectations are calculated by assuming that a site that can vary without changing the amino acid contains a G or C with probability f and an A or T with probability $1 - f$, with the probabilities renormalized so that the e_i add to one.

Correspondence to: Montgomery Slatkin; *email:* slatkin@socrates.berkeley.edu

All twofold redundant amino acids have one codon with a G or C in the third position and the other with A or T. Denoting the codon with G or C in the third position 1 and the other 2, $e_1 = f$ and $e_2 = 1 - f$. For isoleucine, the only threefold redundant amino acid, the codons are AUC ($i = 1$) AUU ($i = 2$), and AUA ($i = 3$), for which $e_1 = f/(2 - f)$, $e_2 = e_3 = (1 - f)/(2 - f)$. For the fourfold amino acids, denote the codons with G and C in the third position 1 and 2 and the others 3 and 4. Then $e_1 = e_2 = f/2$ and $e_3 = e_4 = (1 - f)/2$. There are two types of amino acids with six codons. All codons for serine (UCU [$i = 1$], UCC [$i = 2$], UCA [$i = 3$], UCG [$i = 4$], AGU [$i = 5$], and AGC [$i = 6$] have an A or U in the first codon position and hence $e_1 = e_3 = e_5 = (1 - f)/3$ and $e_2 = e_4 = e_5 = f/3$. The codons for leucine (CUU [$i = 1$], CUC [$i = 2$], CUA [$i = 3$], CUG [$i = 4$], UUA [$i = 5$], and UUG [$i = 6$]) and arginine (CGU [$i = 1$], CGC [$i = 2$], CGA [$i = 3$], CGG [$i = 1$], AGA [$i = 5$], and AGG [$i = 6$]) have a C in the first position of some codons and an A or U in the rest. Hence, for leucine and arginine, $e_1 = e_3 = e_6 = f(1 - f)/(1 + f)$, $e_2 = e_4 = f^2/(1 + f)$, and $e_5 = (1 - f)^2/(1 + f)$.

For each amino acid in each taxon, the data consisted of the numbers of codons, n_i ($i = 1,\ldots k$). The total number of codons for that amino acid in that taxon is $n = \sum_{i=1}^{k} n_i$. The first question is whether there is a significant deviation from the expectation. A χ^2 test with $k - 1$ degrees of freedom and a 5% significance threshold was used to determine whether the n_i differed significantly from ne_i. This χ^2 test differs from that applied by Shields et al. (1988) in that this implementation uses expected values for each

codon that are based on the genomic G+C content, while Shields et al. assume equal usage of synonymous codons. If the null hypothesis that there is no significant difference was not rejected, then there was no significant bias for that amino acid in that taxon, the number of preferred codons was 0, and no further analysis was carried out.

If the null hypothesis was rejected, then a model with one preferred codon was fitted to the data. The codon with the largest value of $p_i = (n_i/n) - e_i$ was identified as the preferred codon. The *single-preference model* assumed that the frequency of usage of the preferred codon is $e_i + b$ and the frequency of usage of the remaining $k - 1$ codons is $e_i - b/(k - 1)$. The sample n_i was assumed to be a multinomial sample of size n. The value of b was estimated by maximizing the likelihood. If the values of p_i indicated that two codons were preferred equally, the single-preference model was not tested because it would necessarily be rejected in favor of the *double-preference model* described next.

If $k > 2$, a model with two preferred codons, the double-preference model, was also fitted to the data. The two codons with the largest values of p_i, were identified and were assigned frequencies $e_i + b_1$ and $e_j + b_2$, and the remaining codons were assigned frequencies $e_i - (b_1 + b_2)/(k-2)$. The values of b_1 and b_2 were estimated by assuming a multinomial distribution and maximizing the likelihood.

Finally, a likelihood-ratio test was used to determine whether the double-preferencc model fit the data significantly better than the single-preference model. The ratio $R = -2\ln(L_2/L_1)$ was assumed to have a χ^2 distribution with 1 degree of freedom, where L_2 is the maximum likelihood for the double-preference model and L_1 is the maximum likelihood for the single-preference model. A 5% significance threshold was used. If R exceeded 3.84, then the double-preference model was accepted, implying that there were two preferred codons. If $R < 3.84$ the single-preference model was accepted, implying that there was one preferred codon. In cases where the double-preference model was accepted, but the b_1 value was 10 times greater than the b_2 value, we chose the single-preference model to describe the data.

This analysis has been implemented in a C++ program that is available at http://ib.berkeley.edu/labs/slatkin/software.html. The program also computes summary statistics such as those described in Wall and Herbeck's paper in this issue (the sum of the bias values and the sum of the χ^2 values) as well as the effective number of codons, N_c (Wright 1990). In addition, the program accommodates the use of alternate genetic codes. Currently, a compiled Linux version of the program is available, and other versions may be made available if sufficient interest exists.