J Mol Evol (1994) 39:533–536

Journal of Molecular Evolution

An Empirical Comparison of Distance Matrix Techniques for Estimating Codon Usage Divergence

David A. Morrison, John Ellis, Alan M. Johnson

School of Biological and Biomedical Sciences, University of Technology, Sydney, PO Box 123, Broadway NSW, 2007, Australia

Received: 15 March 1993 / Accepted: 31 January 1994

Abstract. The quantitative description of codon usage divergence among taxa is a two-step process involving first the calculation of a standardized measure of intertaxon divergence for each possible pair of taxa followed by a summary and visual display of the patterns among the taxa inherent in these measures. Three different measures have been proposed in the literature, and three different summaries have been used. These different techniques are empirically compared using a data set consisting of gene sequences from seven species of Apicomplexa. The results suggest that the manhattan distance measure may be preferable to the use of the chi-square measure, although the separation of amino acid usage and codon usage by the genetic distance produces theoretical advantages. The multidimensional scaling ordination and unweighted pair-group clustering are both successful in displaying the patterns, while the eigenanalysis ordination is not.

Key words: Codon usage — Phylogeny — Distance matrices — Ordination — Clustering — Apicomplexa

Synonymous codons are not used with equal frequencies among taxa, and it is now accepted that there is some close relationship between codon usage divergence and phylogenetic distance (Grantham et al. 1981; Maruyama et al. 1986; Long and Gillespie 1991). A number of distance matrix methods have therefore been proposed for quantitatively describing codon usage divergence among taxa (e.g., Grantham et al. 1981; Long and Gillespie 1991), and in this note we empirically compare these different techniques.

Correspondence to: D.A. Morrison

Long and Gillespie (1991) have also proposed that it may be possible to deduce the phylogeny of the organisms from the patterns of their codon usage divergence among homologous genes. They discuss this idea in detail, and it is certainly worth pursuing the hypothesis that these distance matrix methods may be used to deduce the phylogeny of the organisms concerned. No parsimony, maximum likelihood, or closest tree methods (see Penny et al. 1992) for deducing phylogenies from codon usage data have been reported so far.

The quantitative description of codon usage divergence is a two-step process involving multivariate analysis of the codon frequency data. First, a standardized measure of intertaxon divergence in codon usage is derived, and this is then calculated for each possible pair of taxa. Second, the patterns among the taxa inherent in these measures are summarized and visually displayed. Three different measures have been proposed, and three different summaries have been used. We will compare all three measures and all three summaries (i.e., nine data analyses).

Grantham et al. (1981) used correspondence analysis to quantitatively describe codon usage divergence among unicellular organisms. This technique uses a double-standardized chi-square association measure (Hill 1974) to estimate the codon usage distance among the taxa based on the 61 codon frequencies summed across all of the genes. An eigenanalysis ordination is then used to project the taxa from the multidimensional space defined by the codon frequencies onto a smaller number of dimensions (Hill 1974). The result is then usually visualized as a two-dimensional graph, where the distance between the taxa on the graph represents their differences in codon usage. The biological model un-

derlying the choice of the chi-square association as a measure of codon usage divergence is that differences between taxa are a product of non-independent codon changes between the taxa being compared. This has been the most widely used multivariate method for the quantitative comparison of codon usage among taxa (e.g., Shields and Sharp 1987; Shields et al. 1988; Sharp and Devine 1989; Lloyd and Sharp 1991, 1992).

Ellis et al. (1993) used a somewhat similar approach to that of Grantham et al. (1981). However, they employed a standardized manhattan dissimilarity measure (Faith et al. 1987) to estimate the codon usage distance among the taxa based on the frequency of each codon relative to the other codons for that taxon. They then used an MMDS (metric multidimensional scaling) ordination to produce the two-dimensional graph (Belbin 1991). One of the problems with eigenanalysis ordinations is that the higher-order axes are often polynomial functions of the lower-order axes (Gauch 1982), because the technique is based on a mathematical model requiring orthogonal axes. For this reason, eigenanalysis ordinations might be best applied to data obtained from taxa with a relatively narrow range of variation in codon usage, where this model may apply. MMDS ordinations should avoid these technical problems (Faith et al. 1987; Belbin 1991). The choice of the manhattan distance as a measure of codon usage divergence implies that differences between taxa are a product of independent codon changes between the taxa being compared.

Long and Gillespie (1991) derived their estimate of codon usage distance among vertebrates from Nei's (1972) genetic distance, which is based on a specific population genetic model. The frequency of each codon is calculated relative to the other synonymous codons for that amino acid, thus separating codon usage information from amino acid usage information. A product–moment correlation measure is then used to estimate the similarity between taxa based on these frequencies averaged across all of the 59 codons present, and this is converted to a codon usage distance by taking the negative natural logarithm. An UPGMA (unweighted pair-group method using arithmetic averages) clustering is then used to group the taxa (Sneath and Sokal 1973). The result is then visualized as a dendogram, where the order in which the taxa are joined represents their differences in codon usage. The biological model assumes that individual codon changes are independent.

The relative effectiveness of these multivariate techniques for the analysis of codon usage divergence among taxa needs to be evaluated. This is particularly true since there have been several unfavourable empirical and theoretical critiques of eigenanalysis ordinations in other fields of biology, notably taxonomy (e.g., Rohlf 1972; Pimentel 1981; Hartman 1988) and ecology (e.g., Fasham 1977; Prentice 1977; Kenkel and Orlóci 1986; Minchin 1987).

The relative effectiveness of classification and ordination techniques can be evaluated in terms of (van der Maarel 1979) (1) the structure of the ordination diagram or dendrogram; (2) the separation of clusters in the ordination space, or discrimination between clusters at one particular level of the dendrogram; (3) the biological interpretability of the results; and (4) the agreement of the results with already-established knowledge about the phylogeny of the taxa involved. Our empirical evaluation will concentrate on the last of these criteria.

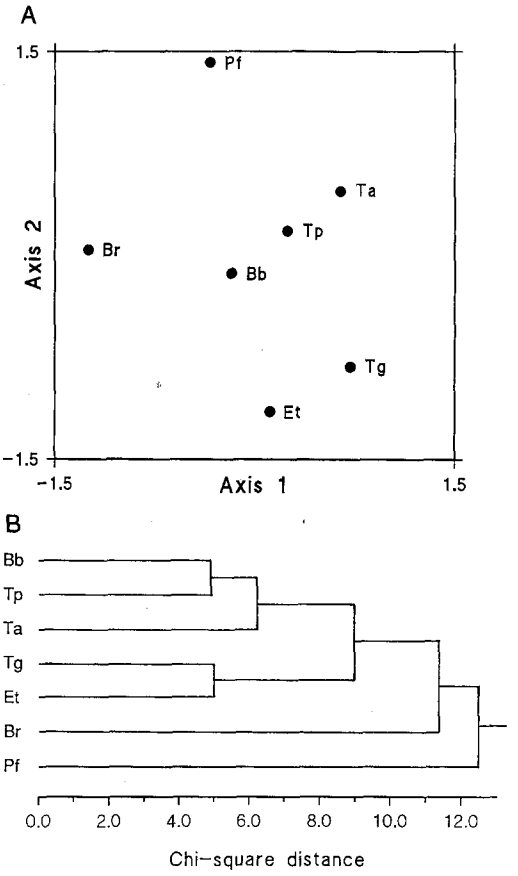

Fig. 1. MMDS ordination (**a**) and UPGMA clustering (**b**) of the Apicomplexa data using the chi-square association as an estimate of codon usage divergence among the taxa. Taxon codes are *Babesia bovis* (*Bb*), *Babesia rodhaini* (*Br*), *Eimeria tenella* (*Et*), *Plasmodium falciparum* (*Pf*), *Theileria annulata* (*Ta*), *Theileria parva* (*Tp*), *Toxoplasma gondii* (*Tg*).

The gene sequences used for our empirical comparison are described by Ellis et al. (1993). They are a set of nonhomologous sequences from seven species of Apicomplexa, a group of parasitic protozoa. Long and Gillespie (1991) point out that comparisons of nonhomologous sequences may confound codon choice with amino acid choice, i.e., that differences in amino acid usage among the sequences will also contribute to the estimates of codon usage distances. However, we can assess the extent of this potential problem by comparing our results with the conclusions drawn from other data sets for the same taxa.

We applied the chi-square distance (as suggested by

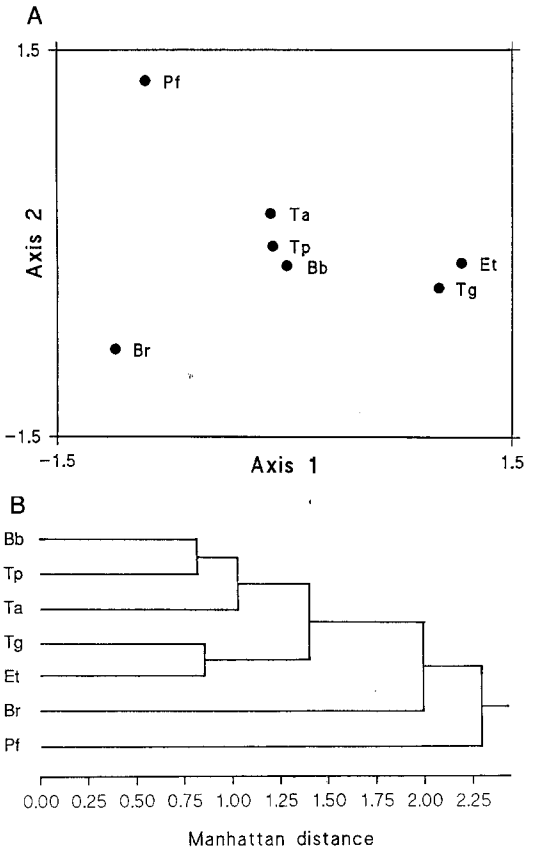

Fig. 2. MMDS ordination **(a)** and UPGMA clustering **(b)** of the Apicomplexa data using the manhattan dissimilarity as an estimate of codon usage divergence among the taxa. Taxon codes are as in Fig. 1.

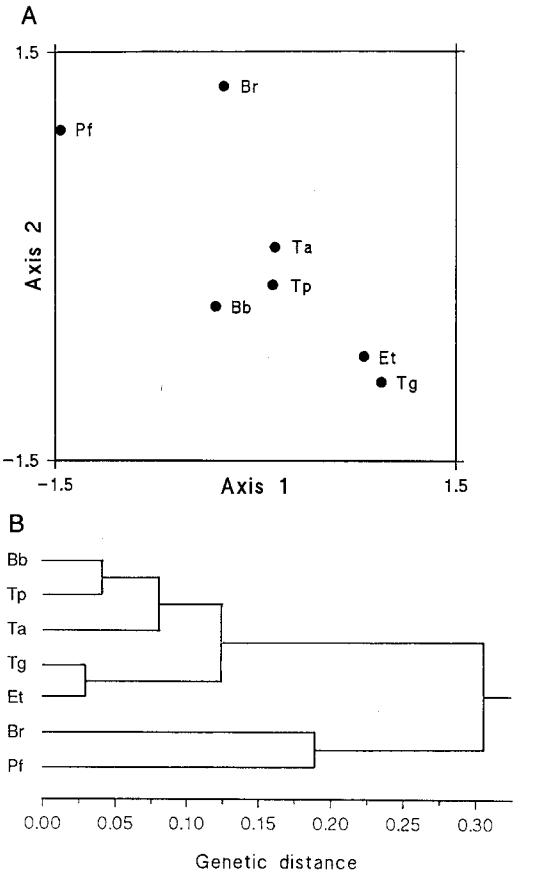

Fig. 3. MMDS ordination **(a)** and UPGMA clustering **(b)** of the Apicomplexa data using the genetic distance as an estimate of codon usage divergence among the taxa. Taxon codes are as in Fig. 1.

Grantham et al. 1981), the manhattan distance (as suggested by Ellis et al. 1993), and the genetic distance (as suggested by Long and Gillespie 1991) to the Apicomplexa data. We then used the UPGMA clustering to group the data (as used by Long and Gillespie 1991), and the MMDS ordination (as used by Ellis et al. 1993). The eigenanalysis ordination (as used by Grantham et al. 1981) was also calculated for comparison. All analyses used the PATN multivariate data analysis package (Belbin 1989).

The results of the MMDS ordination and UPGMA clustering analyses of the chi-square (Fig. 1), manhattan (Fig. 2), and genetic distance (Fig. 3) estimates of codon usage divergence are in general agreement with the established phylogenetic relationships among the taxa. These relationships include those indicated by aspects of both morphology (Levine 1985) and genetics (Barta et al. 1991; Johnson et al. 1991; Ellis et al. 1992, 1993). This result implies that the MMDS ordination and UPGMA clustering analyses are useful distance matrix techniques for analyzing codon usage divergence.

This result also confirms the hypothesis discussed by Long and Gillespie (1991) that some distance matrix methods based on codon usage divergence may be a useful means of detecting the phylogenetic relationships among taxa. In particular, our results extend the idea to cover prokaryotic organisms, in addition to the vertebrates specifically discussed by Long and Gillespie (1991).

The MMDS ordination and UPGMA clustering analyses within each distance measure also agree closely with each other. This suggests that these two summarization techniques do not differ in their ability to detect and display the patterns of codon usage in the data, and either technique could be used to estimate phylogeny. However, the three eigenanalysis ordinations do not produce results that agree with either the equivalent MMDS ordination or the equivalent UPGMA clustering, implying that this ordination techniques is less successful, presumably due to its restrictive mathematical model. Therefore, we have not presented these results. These conclusions agree with those of the other published studies that have been critical of eigenanalysis ordinations. (See James and McCulloch 1990.)

The three distance measures do disagree in some of the details of the relationships among the taxa. First, the genetic and manhattan distance measures group the taxa quite tightly, thus emphasizing the phylogenetic relationship among the taxa more than does the chi-square measure. Second, the distance measures disagree about the placement of *Plasmodium falciparum*—they all agree that it is only distantly related to the other taxa,

but the genetic distance groups this taxon with *Babesia rodhaini* whereas the chi-square and manhattan distances place it in a group on its own. This difference may be a result of the attempt by the genetic distance to separate the codon usage information from the amino acid usage information, which the other two measures do not do. The other morphological and genetic data sets indicate that it is unlikely that *P. falciparum* is phylogenetically more closely related to *B. rodhaini* than it is to the other taxa, suggesting that it is the manhattan distance that is producing groupings that approximate the most likely phylogeny.

We thus conclude that the manhattan distance measure may be preferable to the alternative distance measures as an estimate of divergence among taxa in codon usage, although the separation of amino acid usage and codon usage by the genetic distance produces theoretical advantages as well. The MMDS ordination and UPGMA clustering are both successful in displaying the patterns of divergence, while the eigenanalysis ordination is not.

References

Barta JR, Jenkins MC, Danforth HD (1991) Evolutionary relationships of avian *Eimeria* species among other apicomplexan protozoa based on partial small subunit ribosomal RNA sequences: monophyly of the Apicomplexa is supported. Mol Biol Evol 8:345–355

Belbin L (1989) PATN—pattern analysis package technical reference. CSIRO, Canberra

Belbin L (1991) Semi-strong hybrid scaling, a new ordination algorithm. J Veg Science 2:491–496

Ellis J, Hefford C, Baverstock PR, Dalrymple BP, Johnson AM (1992) Ribosomal DNA sequence comparison of *Babesia* and *Theileria*. Mol Biochem Parasitol 54:87–96

Ellis J, Griffin H, Morrison D, Johnson AM (1993) Analysis of dinucleotide frequency and codon usage in the phylum Apicomplexa. Gene 126:163–170

Faith DP, Minchin PR, Belbin L (1987) Compositional dissimilarity as a robust measure of ecological distance. Vegetatio 69:57–68

Fasham MJR (1977) A comparison of nonmetric multidimensional scaling, principal components and reciprocal averaging for the ordination of simulated coenoclines and coenoplanes. Ecology 58:551–561

Gauch HG (1982) Multivariate analysis in community ecology. Cambridge University Press, New York

Grantham R, Gautier C, Gouy M, Jacobzone M, Mercier R (1981) Codon catalog usage is a genome strategy modulated for gene expressivity. Nucleic Acids Res 9:r43–r74

Hartman SE (1988) Evaluation of some alternative procedures used in numerical systematics. Syst Zool 37:1–18

Hill MO (1974) Correspondence analysis: a neglected multivariate method. J R Stat Soc 23:340–354

James FC, McCulloch CE (1990) Multivariate analysis in ecology and systematics: panacea or Pandora's box? Ann Rev Ecol Syst 21:129–166

Johnson AM, Fielke R, Ellis J, O'Donoghue PJ, Baverstock PR (1991) The phylogenetic relationships of the genus *Eimeria* based on comparison of partial sequences of 18S rRNA. Syst Parasitol 18:1–8

Kenkel NC, Orlóci L (1986) Applying metric and nonmetric multidimensional scaling techniques to ecological studies: some new results. Ecology 67:919–928

Levine ND (1985) Phylum II. Apicomplexa levine, 1970. In: Lee JJ, Hunter SH, Bovee EC (eds) Illustrated guide to the protozoa. Society Protozoologists, Kansas, pp 322–374

Lloyd AT, Sharp PM (1991) Codon usage in *Aspergillus nidulans*. Mol Gen Genet 230:288–294

Lloyd AT, Sharp PM (1992) Evolution of codon usage patterns: the extent and nature of divergence between *Candida albicans* and *Saccharomyces cerevisiae*. Nucleic Acids Res 20:5289–5295

Long M, Gillespie JH (1991) Codon usage divergence of homologous vertebrate genes and codon usage clock. J Mol Evol 32:6–15

Maruyama T, Gojobori T, Aota S-I, Ikemura T (1986) Codon usage tabulated from the GenBank genetic sequence data. Nucleic Acids Res 14:r151–r197

Minchin P (1987) An evaluation of the relative robustness of techniques for ecological ordination. Vegetatio 69:89–107

Nei M (1972) Genetic distance between populations. Am Nat 106:283–292

Penny D, Hendy MD, Steel MA (1992) Progress with methods for constructing evolutionary trees. Trends Ecol Evol 7:73–79

Pimental RA (1981) A comparative study of data and ordination techniques based on a hybrid swarm of sand verbenas (*Abronia* Juss.). Syst Zool 30:250–267

Prentice IC (1977) Non-metric ordination methods in ecology. J Ecol 65:85–94

Rohlf FJ (1972) An empirical comparison of three ordination techniques in numerical taxonomy. Syst Zool 21:271–280

Sharp PM, Devine KM (1989) Codon usage and gene expression level in *Dictyostelium discoideum:* highly expressed genes do 'prefer' optimal codons. Nucleic Acids Res 17:5029–5039

Shields DC, Sharp PM (1987) Synonymous codon usage in *Bacillus subtilis* reflects both translational selection and mutational biases. Nucleic Acids Res 15:8023–8040

Shields DC, Sharp PM, Higgins DG, Wright F (1988) "Silent" sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. Mol Biol Evol 5:704–716

Sneath PHA, Sokal RR (1973) Numerical taxonomy. Freeman, San Francisco

van der Maarel E (1970) Transformation of cover-abundance values in phytosociology and its effects on community similarity. Vegetatio 39:97–114