JOURNAL OF MOLECULAR EVOLUTION

Compositional Bias May Affect Both DNA-Based and Protein-Based Phylogenetic Reconstructions

Peter G. Foster,* Donal A. Hickey

Department of Biology, University of Ottawa, Ottawa, Ontario, Canada K1N 6N5

Received: 19 February 1998 / Accepted: 28 August 1998

Abstract. It is now well-established that compositional bias in DNA sequences can adversely affect phylogenetic analysis based on those sequences. Phylogenetic analyses based on protein sequences are generally considered to be more reliable than those derived from the corresponding DNA sequences because it is believed that the use of encoded protein sequences circumvents the problems caused by nucleotide compositional biases in the DNA sequences. There exists, however, a correlation between AT/GC bias at the nucleotide level and content of AT- and GC-rich codons and their corresponding amino acids. Consequently, protein sequences can also be affected secondarily by nucleotide compositional bias. Here, we report that DNA bias not only may affect phylogenetic analysis based on DNA sequences, but also drives a protein bias which may affect analyses based on protein sequences. We present a striking example where common phylogenetic tools fail to recover the correct tree from complete animal mitochondrial protein-coding sequences. The data set is very extensive, containing several thousand sites per sequence, and the incorrect phylogenetic trees are statistically very well supported. Additionally, neither the use of the LogDet/paralinear transform nor removal of positions in the protein alignment with AT- or GC-rich codons allowed recovery of the correct tree. Two taxa with a large compositional bias continually group together in these analyses, despite a lack of close biological relatedness. We conclude that even protein-based phylogenetic trees may be misleading, and we advise caution in phylogenetic reconstruction using protein sequences, especially those that are compositionally biased.

Key words: G + C content — Composition bias — Phylogenetic analysis — Mitochondrial genes

Introduction

Hasegawa and Hashimoto (1993) pointed out that phylogenetic analyses based on rRNA genes could be unreliable due to extreme AT or GC nucleotide bias in the rRNA genes of some taxa. They suggested that the inferred amino acid sequences of encoded proteins provide more reliable phylogenies. Many molecular evolutionists now agree that protein sequences are relatively free from the effects of nucleotide bias (Loomis and Smith 1990; Lockhart et al. 1992). This view is based on the assumption that, while DNA may be driven to extremes of AT or GC bias by directional mutation pressure, the protein composition remains constant, due to the greater functional constraints on the protein sequence. Contrary to this assumption, we have recently shown (Foster et al. 1997) that amino acid sequences can be compositionally biased in a manner that parallels the nucleotide composition of the codons. For instance, we showed that those animal mitochondrial genes which are most AT-rich at the DNA level tend to be rich in those amino acids which are encoded by AT-rich codons, ie, codons with either A or T in the first and second codon position; this set in-

* *Present address:* Laboratory of Molecular Systematics, National Museum of Natural History, Smithsonian Institution, Washington, DC 20560
Correspondence to: D.A. Hickey; *e-mail:* dhickey@uottawa.ca

cludes the codons for phenylalanine (F), tyrosine (Y), methionine (M), isoleucine (I), asparagine (N), and lysine (K). These same proteins are correspondingly poor in amino acids coded for by GC-rich codons: glycine (G), alanine (A), arginine (R), and proline (P). This effect is not limited to animal mitochondrial genes; it has been reported for a wide range of genes and genomes (Sueoka 1961; Andersson and Sharp 1996; Collins and Jukes 1993; Porter 1995; Jukes and Bhushan 1986; Jermiin et al. 1994; D'Onofrio et al. 1991), although there are some genes that appear to be immune to this effect (Hashimoto et al. 1994, 1995). The recent publication of the complete genome sequence of *Mycobacterium tuberculosis* (Cole et al. 1998) provides an excellent example of the correlation between GC bias and amino acid compositional bias.

Here we ask if sequences with similarly-biased compositions will tend to be grouped together in molecular phylogenetic analysis, even if they do not share a recent common ancestor. We performed phylogenetic analyses of the protein coding sequences of several mitochondria, which included taxa with varying amounts of both DNA and amino acid composition bias. Maximum-likelihood analysis of the DNA sequences failed to find the correct tree. Our available arsenal of phylogenetic tools for phylogenetic analysis of protein sequences, including distance and parsimony methods, maximum-likelihood, and LogDet distance correction, also failed to recover the correct tree. Two taxa with a large compositional bias, the honeybee and the nematode, continually grouped together in these analyses, despite lack of close biological relatedness.

Results

We have chosen mitochondrial genes from animal species for which the entire mitochondrial genome has been sequenced and we have used the concatenated protein-coding sequences for our analyses. The species cover a broad phylogenetic range within the metazoa and the pattern of their true phylogenetic divergences has consensus (Fig. 1A) (Maddison and Maddison 1997; but see Aguinaldo et al. 1997, who place nematodes at the base of the arthropod clade). They include taxa that are very AT-rich, and show the predicted bias in amino acid composition (Foster et al. 1997; Table 1). That is, the taxa which are AT-rich at the DNA level are correspondingly high in AT-rich codons, coding for the amino acids F, Y, M, I, N, and K, and poor in CG-rich codons, coding for the amino acids G, A, R, and P. The goal of our study was to test in phylogenetic reconstruction if the signal from compositional bias can override the signal from common ancestry. This would be indicated by two compositionally biased taxa grouping together without being true sister taxa.

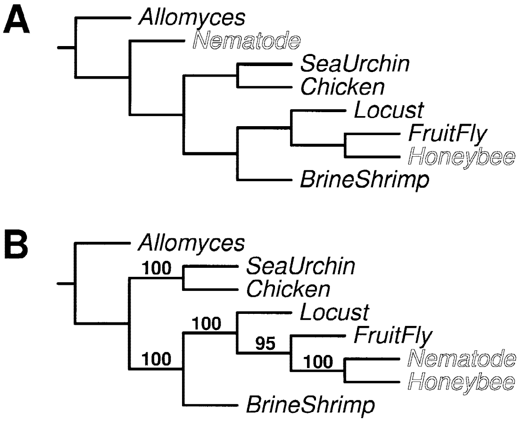

Fig. 1. **(A)** Consensus biological tree. **(B)** Maximum-likelihood tree based on DNA sequences. *Numbers* indicate bootstrap support.

Table 1. Composition bias in mitochondrial protein-coding genes[a]

	Nucleotide (% G + C)	Protein % FYMINK	% GARP	FYMINK/ GARP
Honeybee[b]	17.2	49.1	9.5	5.2
Nematode	24.8	38.8	12.3	3.2
Locust	26.3	38.4	15.4	2.5
Fruit fly	23.8	35.6	16.2	2.2
Brine Shrimp	36.3	32.0	17.2	1.9
Allomyces	32.8	31.4	20.4	1.5
Sea Urchin	41.0	28.9	21.1	1.4
Chicken	47.3	26.3	21.9	1.2

[a] Compositional bias of the taxa used, ordered by decreasing FYMINK/GARP. In the DNA alignment, the proportion of G + C is shown, and in the protein alignment the proportion of AT-rich FYMINK amino acids (Phe, Tyr, Met, Ile, Asn, and Lys), GC-rich GARP amino acids (Gly, Ala, Arg, and Pro), and the ratio between them are shown.
[b] Genbank accession numbers L06178, X54252, X80245, X03240, X69067, U41288, J04815, and X52392, respectively.

Protein sequences of the 12 protein-coding genes common to all the taxa were aligned individually using clustalw (Thompson et al. 1994). The ragged ends of the individual alignments were trimmed, and then the alignments were concatenated to make an alignment 3713 amino acids in length. DNA sequences were then aligned to this protein alignment to make a DNA alignment 11139 nucleotides in length. Mitochondrial sequences from the fungus *Allomyces macrogynus* were used as an outgroup to the animal mitochondrial sequences.

Our strategy was to first perform a phylogenetic analysis using DNA sequences only. After we observed the predicted misclustering of the nematode and the honeybee (presumably due to their common nucleotide bias), we then asked if this problem could be circumvented by using the encoded protein sequences rather than the DNA sequences themselves. As shown below, the grouping of the nematode with the honeybee persisted even when the amino acid sequences were used instead of the DNA sequences.

Phylogenetic Analysis of DNA Sequences

The maximum-likelihood analysis was used, using PAUP*, v4.0.0d64 (Swofford, 1998). The model for analysis was chosen by the following method. First, pairwise maximum likelihood distances were calculated using the HKY model (Hasegawa et al. 1985) and used to make a neighbor-joining tree. Using this tree, various models were evaluated by the likelihood ratio test. The best model was the general time reversible model with gamma-distributed rates (four categories in a discrete gamma approximation) allowing invariant sites (Swofford et al., 1996). Having chosen this model, using the parameters derived from the neighbor-joining tree, a preliminary heuristic search for the most likely tree was made. Using parameters derived from the best tree found in this search, a final search was made using the branch and bound strategy, which confirmed the most likely tree (Fig. 1B). Bootstrap analysis used these parameters with heuristic searches at each bootstrap resampling.

In the most likely tree (Fig. 1B) the nematode groups with the honeybee with high statistical confidence, yet this tree cannot be correct. We implicate the shared compositional bias as being a major factor in why these unrelated taxa group together. This cannot be the complete explanation, as the fruit fly is biologically more closely related to the honeybee than is the nematode and is slightly more AT-rich than the nematode (Table 1). Among the several signals in these sequences, the signal due to compositional bias appears to be one of the signals that overwhelm the signal due to common ancestry in this case. If the same analysis is done without the honeybee sequence, the nematode appears between the shrimp and the insects (in a tree which is otherwise congruent to the biological tree shown in Fig. 1A), and if the analysis is done without the nematode, the Fig. 1A tree (minus the nematode) is obtained.

The remainder of this study addresses the question of whether a parallel problem exists in phylogenetic reconstruction based on protein sequences.

Distance, Parsimony, and Maximum-Likelihood Methods Fail to Find the Correct Tree Based on Protein Sequences

We used the amino acid alignment described above to construct phylogenetic trees. In addition to maximum likelihood, we also used both the distance-based neighbor-joining method, and the method of maximum parsimony (Felsenstein, 1993). These results (Fig. 2) show that, although the honeybee and nematode are widely separated in evolutionary time (Fig. 1A), they are erroneously grouped together in all three of the computed phylogenetic trees. Despite being incorrect, these computed trees are very well supported as indicated by the high bootstrap and quartet puzzling values.

If we perform the maximum likelihood analysis excluding both the honeybee and the nematode, the correct (Fig. 1A) tree is obtained. If we include the honeybee but not the nematode, the maximum likelihood tree incorrectly places the honeybee between the shrimp and the other two insects. If we include the nematode but not the honeybee, the nematode is placed at the base of the arthropod clade in an otherwise correct tree.

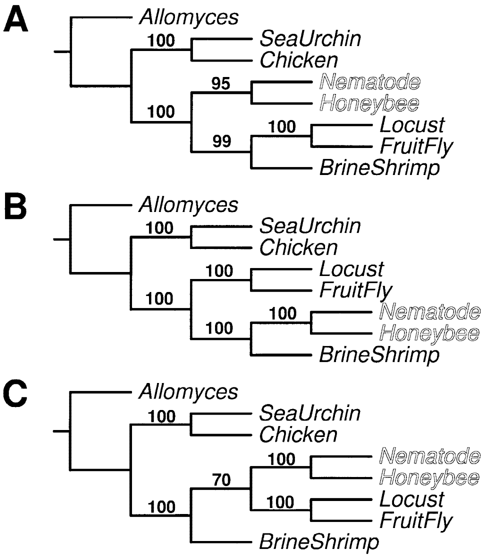

Fig. 2. Protein-based phylogenetic analyses. **(A)** Protdist neighbor-joining tree. **(B)** Tree from protpars. For both A and B, the numbers indicate bootstrap support. **(C)** Maximum-likelihood tree from both protml (using an exhaustive search with the JTT-F model) and puzzle (using the mtRev model, with a mixed rate model allowing invariant sites and eight categories of gamma-distributed rates, and with frequencies, rates, and invariant sites proportion estimated from the data). For this tree *numbers* indicate quartet support.

Erroneous trees were also obtained when individual mitochondrial genes from these taxa were examined. The protein sequences of the *cob, cox1,* and *nad5* genes were analyzed by all three methods as described above. Trees of various configurations, often one of the trees shown in Fig. 2, were obtained. In no case was the correct biological tree obtained with these methods, and in all but one case the nematode grouped with the honeybee (the exception being analysis of *cox1* sequences using the maximum-likelihood program puzzle, which placed both the nematode and the honeybee in the insects).

The effect of a change of outgroup was examined, again using the full set of sequences common to all the mitochondria examined. When the mitochondrial sequences of the liverwort (*Marchantia polymorpha;* accession number M68929; 32.4% FYMINK, 21.5% GARP, FYMINK/GARP = 1.5) were substituted for those of *Allomyces* as the outgroup to the analysis, identical ingroup topologies were obtained, the same as Figs. 2A and B for neighbor-joining and maximum parsimony, respectively. Using protml and puzzle as described in the legend to Fig. 2, the maximum-likelihood trees were as shown in Figs. 2B and 2C, respectively.

LogDet/Paralinear Transform

Phylogenetic analyses commonly assume that trees are homogeneous, that is, that the same model applies throughout the tree, and that the sequence composition is stationary. The LogDet/paralinear transform is a method of calculating a distance matrix which is able to recover the correct tree when sequences evolve under nonhomogeneous, nonstationary models (Lockhart et al. 1994; Lake 1994). There are two forms of the LogDet transform, which are given in Eqs. 1 and 3 of Lockhart et al. (1994). Equation 3 is equivalent, with scaling, to the paralinear distance (Lake 1994; Swofford et al. 1996).

Care needs to be taken in the choice of which sites of an alignment to include in the calculation of the LogDet/paralinear distance. Inclusion of invariant sites in the distance calculation tends to misestimate the amount of change (Lockhart et al. 1994, 1996). Additionally, sites which vary a great deal are problematic because of saturation. It has been shown to be useful to exclude both of these extremes by using only parsimony sites (Lockhart et al. 1994). Another area where care is required in these calculations stems from the use of the logarithm of the determinant of the matrix of transitions between the two sequences between which the distance is being calculated. The calculation can result in a negative determinant, for which the logarithm is undefined. The interpretation in this case would be that there is such a large divergence between the two taxa that the sequences are effectively random. The distance between those taxa is then arbitrarily large. In order to tree the distance matrix, using for example the neighbor-joining algorithm, one needs to choose an arbitrarily large number as the distance between these problem taxa. For example, the program PAUP*, Version 4.0, sets the values of these undefined distances at twice the distance of the largest defined distance in the distance matrix. However, the choice of this distance affects the tree topology, and so caution is needed in interpreting such trees.

We calculated LogDet distances using both Eq. 1 and Eq. 3 (Lockhart et al. 1994), using all 3713 sites or using only the 1715 parsimony sites. When using parsimony sites there were nine pairwise comparisons which had negative determinants. We set these to 1.1×, 2×, and 10× the largest defined distance. The resulting distance matrices were then analyzed using the neighbor-joining method. When all sites were used, both Eq. 1 and Eq. 3 resulted in a tree of the same topology as the tree shown in Fig. 2A. LogDet calculations based on parsimony sites resulted in trees of various other configurations. In no case was the correct biological tree obtained, and so it appears that this data set is intractable to correction by the LogDet/paralinear transform.

Removal of Amino Acid Groups

We can speculate that the honeybee and nematode mitochondrial proteins have independently become ‘‘FYMINK-rich’’ at the amino acid level, due to AT pressure at the nucleotide level, and that many of these FYMINK amino acids happen to be at homologous sites in the two sequences. Phylogenetic algorithms then mistake this correspondence as relatedness due to recent common ancestry, and consequently group the sequences together in the inferred tree. We tested whether this was correctable by simply removing AT- or GT-rich amino acids. Entire columns were removed, thereby preserving the alignment.

We first removed all sites in the alignment which contained any of the FYMINK amino acids. The remainder was analyzed with protdist/neighbor-joining and with protpars (Fig. 3A). The distance and parsimony tree topologies for this shortened alignment of 993 positions are the same as that for the entire alignment of 3713 positions, although some bootstrap values are lower. The nematode and honeybee still group together. Similar results were obtained when, in addition to the FYMINK set of amino acids, leucine was also removed (Fig. 3B). Recall that in all these taxa leucine has two codon families, one of which is AT-rich, while the other is AT-neutral. When the 1422 positions which contained any of the GARP amino acids were removed, we obtained the trees shown in Fig. 3C. Again, the honeybee and nematode are found together or nearby in the tree, and the bootstrap values are somewhat smaller. We then removed all the positions in the alignment which contained any of FLYMINK (including leucine) or GARP amino acids (Fig. 3D). This perturbed the resulting trees somewhat more, tending to separate the honeybee and nematode. This short alignment of only 360 amino acids did not result in the correct tree, and the bootstrap values were low. The maximum-likelihood trees from these subset alignments, found as described in the legend to Fig. 2, were of various configurations, usually different from those from neighbor-joining and parsimony. In half of the maximum-likelihood trees the nematode grouped with the honeybee, and all differed from the biological tree as shown in Fig. 1A.

Discussion

A number of previous studies have shown that biased nucleotide composition can affect phylogenetic reconstruction based on DNA sequences. Here we have provided another example showing that compositional bias in DNA sequences can adversely affect phylogenetic analysis. To this observation, however, we add that even protein sequence analysis can be so affected. The neighbor-joining, maximum-parsimony, and maximum-likelihood methods all failed to reconstruct the correct phylogeny from entire mitochondrial protein sequences. Not only did they fail, but they indicated incorrect trees with high statistical confidence. No fault can be found

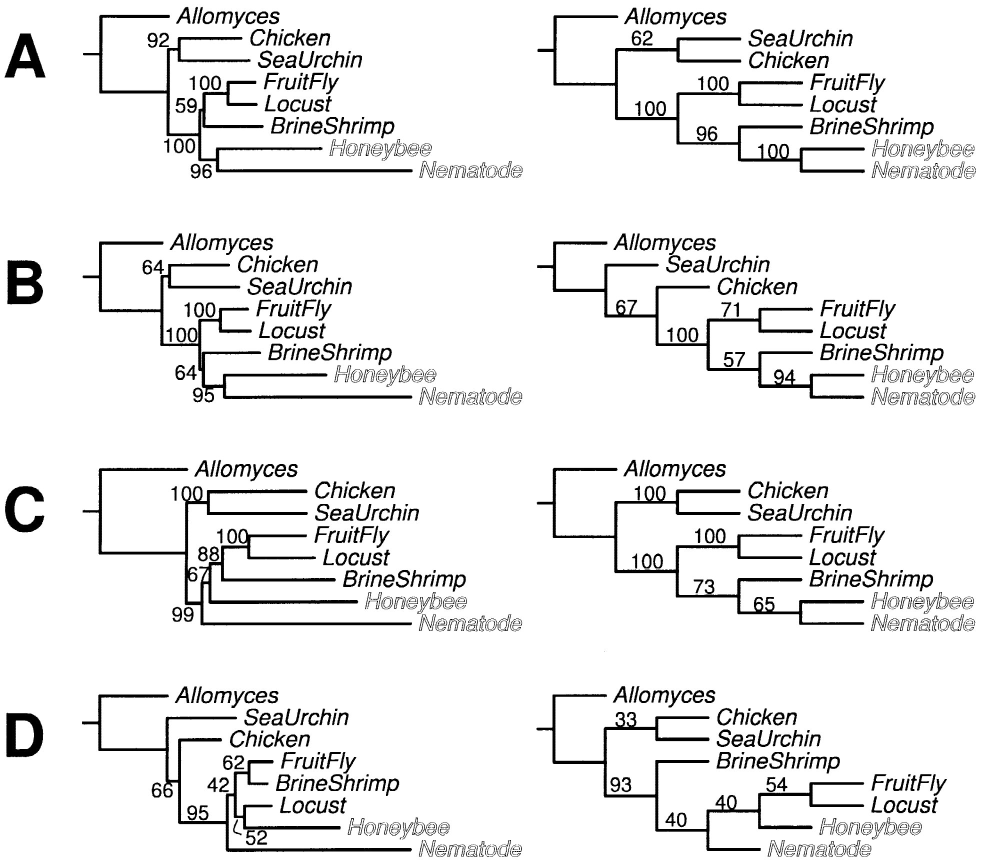

Fig. 3. Removal of amino acids from the alignment before analysis with the distance-based protdist/neighbor-joining on the *left* and by maximum parsimony using protpars on the *right*. Neighbor-joining branch lengths are meaningful, but the branch lengths of the parsimony analysis have been equalized. *Numbers* are bootstrap values, the percentage of 100 bootstraps. The original alignment is 3713 positions. **(A)** FYMINK removed, leaving 993 positions. **(B)** FLYMINK removed, leaving 806 positions. **(C)** GARP removed, leaving 2291 positions. **(D)** Both FLYMINK and GARP removed, leaving 360 positions.

with the choice of sequences, as mitochondrial sequences are commonly used in phylogenetics, and the use of the entire genome is considered especially reliable (Russo et al. 1996). In addition, the LogDet/paralinear transform did not allow reconstruction of the correct tree, even when only parsimony sites were used (Lockhart et al. 1994; Lake 1994). Considering the size of the sequence set, the analysis appears to be inconsistent (Hillis et al. 1994), meaning that the analysis does not converge to the correct answer with increasing amounts of data. Nei (1996) describes a detailed phylogenetic problem in using mitochondrial total amino acid sequences. When 11 vertebrate species were examined, a correct and well-supported phylogeny was obtained. However when lamprey and sea urchin sequences were incorporated, an incorrect phylogeny was obtained with high bootstrap values using several tree-building methods. The reason for this was not clear. Naylor and Brown (1998) describe a similar problem with phylogenetic reconstruction based on mitochondrial sequences. Using both DNA and protein sequences, they obtained incorrect trees with high bootstrap support. In their opinion the incorrect trees resulted in part from convergent base compositional similarities. Consistent with our analysis at the protein level, they were not able to recover correct trees using the LogDet/paralinear transform at the DNA level. However, they were able to recover the correct tree with high support using protein sequences from a subset of mitochondrial genes, suggesting that compositional bias of their data was not as pronounced as that observed in our data.

We tested the possibility that similarly biased taxa tended to group together (''attract'') solely because of an increase in AT-rich codons, or a decrease in GC-rich codons, by removal of positions in the alignment where these amino acid groups were found. Again, the correct tree was not obtained. It is an oversimplification to say that the bias resides only in the GARP and FYMINK(L) groups of amino acids. An AT/GC-neutral ancestor sequence which becomes AT-rich over time will of course not convert its GARP residues solely to FYMINK(L) residues but, rather, will tend to lose GARP amino acids to more or less any other amino acid. Similarly the increase in FYMINK occurs from positions which could have been any other amino acid, not just GARP. Although we see the difference in frequencies in GARP and FYMINK, and not in the other amino acids, the bias will have an effect at those other amino acids as well.

That compositional bias can affect phylogenetic reconstruction has been shown here, but it is difficult to predict what will happen to an analysis based only on compositional biases. Thus in the analyses of protein sequences, the two taxa with the greatest amount of FYMINK and the least amount of GARP (honeybee and nematode) grouped together. However in analysis of the DNA sequences, the two taxa with the greatest AT bias (honeybee and fruit fly) did not group together. We can suppose that there are several signals in the sequences, some of which come from common ancestry, some from amino acid composition bias, and other signals. It appears from this study that the signal due to composition bias can pervade the sequences and overwhelm or hide the signals due to common ancestry.

One possible approach to understanding why the honeybee and nematode group together is to perform the phylogenetic reconstruction in the absence of one or other of these taxa. When this was done, in both cases the remaining taxon had a long branch. When we deleted the honeybee sequence, using DNA sequences we found that the nematode fell basal to the insects, and using protein sequences we found that the nematode fell basal to the arthropods (similar to that found by Aguinaldo et al. 1997), while the consensus view places the nematodes branch before the protostome–deuterostome split (Maddison and Maddison, 1997). This is consistent with the view that the insect sequences, which are FYMINK-rich as a whole, ‘‘attract’’ the FYMINK-rich nematode sequence. When, in turn, we omitted the nematode sequence, we found that the honeybee branches out before the other insects. This suggests that the extremely biased composition of the honeybee proteins exaggerates the distance between these sequences and those of related insects. When both the nematode and honeybee are included we have the problem of long branches compounded with correlated compositional changes. The combination of these two effects is enough to confound all phylogenetic reconstruction methods that we tried. Even maximum-likelihood, which is relatively immune to long branch effects, was unable to resolve these data correctly.

We conclude that phylogenetic trees based on amino acid sequences can indeed be misleading because they are subject to the effects of compositional biases. In the case we have described here, the incorrect result is very well supported statistically. This is because we have used a large data set (several thousand amino acids from each taxon) and because we deliberately chose an example where the differences in amino acid composition are pronounced. More subtle biases will cause similar problems, however, in cases where the real phylogenetic distinctions are more difficult, such as the analysis of very ancient divergences. It will be of special interest to look for the possible effects of compositional bias in those protein-based phylogenies which have been the subject of much recent debate (Golding and Gupta 1995; Doolittle et al. 1996; D’Erchia et al. 1996). For instance, in one of these studies (D’Erchia et al. 1996), the molecular phylogeny was deemed to be highly reliable based on the consistency of the results obtained by different methodological approaches, the large number of sites included in the analysis, and the very significant bootstrap values obtained. In the example we have given here all of these criteria are also met, but for a molecular phylogeny that is obviously wrong. This indicates that, despite the power of molecular phylogenetic inference, caution is warranted in the interpretation of all molecular phylogenies. This caution will be especially relevant to the phylogenetic analysis of whole genomes that are subjected to correlated DNA and amino acid biases, such as the recently-sequenced *Mycobacterium tuberculosis* genome (Cole et al. 1998).

Acknowledgments. This research was supported by a grant from the Natural Sciences and Engineering Research Council of Canada to D.A.H. We thank Brian Golding and Lars S. Jermiin for their comments on the manuscript and Peter Lockhart for his comments on the LogDet transform. We thank David Swofford for the use of a prerelease version of PAUP*.

References

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489–93

Andersson SGE, Sharp PM (1996) Codon usage and base composition in rickettsia prowazekii. J Mol Evol 42:525–536

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Barrell BG, et al. (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393:537–544

Collins DW, Jukes TH (1993) Relationship between G + C in silent sites of codons and amino acid composition of human proteins. J Mol Evol 36:201–213

D’Erchia AM, Gissi C, Pesole G, Saccone C, Arnason U (1996) The guinea-pig is not a rodent. Nature 381:597–600

D’Onofrio G, Mouchiroud D, Aissani B, Gautier C, Bernardi G (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. J Mol Evol 32:504–510

Doolittle RF, Feng DF, Tsang S, Cho G, Little E (1996) Determining divergence times of the major kingdoms of living organisms with a protein clock. Science 271:470–477

Felsenstein J (1993) PHYLIP (phylogeny inference package), Version 3.5c. Distributed by the author. Department of Genetics, University of Washington, Seattle

Foster PG, Jermiin LS, Hickey DA (1997) Nucleotide composition bias affects amino acid content in proteins coded by animal mitochondria. J Mol Evol 44:282–288

Golding GB, Gupta RS (1995) Protein-based phylogenies support a chimeric origin for the eukaryotic genome. Mol Biol Evol 12:1–6

Hasegawa M, Hashimoto T (1993) Ribosomal RNA trees misleading? Nature 361:23

Hasegawa M, Kishino H, Yano T (1985) Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol 22:160–174

Hashimoto T, Nakamura Y, Nakamura F, Shirakura T, Adachi J, Goto N, Okamoto K, Hasegawa M (1994) Protein phylogeny gives a robust estimation for early divergences of eukaryotes: Phylogenetic place of a mitochondria-lacking protozoan, Giardia lamblia. Mol Biol Evol 11:65–71

Hashimoto T, Nakamura Y, Kamaishi T, Nakamura F, Adachi J, Okamoto K, Hasegawa M (1995) Phylogenetic place of mitochondrion-lacking protozoan, Giardia lamblia, inferred from amino acid sequences of elongation factor 2. Mol Biol Evol 12:782–793

Hillis DM, Huelsenbeck JP, Swofford DL (1994) Hobgoblin of phylogenetics? Nature 369:363–364

Jermiin LS, Graur D, Lowe RM, Crozier RH (1994) Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome b genes. J Mol Evol 39:160–173

Jukes TH, Bhushan V (1986) Silent nucleotide substitutions and G + C content of some mitochondrial and bacterial genes [published erratum appears in J Mol Evol 1987;24(4):380]. J Mol Evol 24:39–44

Lake JA (1994) Reconstructing evolutionary trees from DNA and protein sequences: paralinear distances. Proc Natl Acad Sci USA 91: 1455–1459

Lockhart PJ, Howe CJ, Bryant DA, Beanland TJ, Larkum AW (1992) Substitutional bias confounds inference of cyanelle origins from sequence data. J Mol Evol 34:153–162

Lockhart PJ, Steel MJ, Hendy MD, Penny D (1994) Recovering evolutionary trees under a more realistic model of sequence evolution. Mol Biol Evol 11:605–612

Lockhart PJ, Larkum AW, Steel M, Waddell PJ, Penny D (1996) Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc Natl Acad Sci USA 93: 1930–1934

Loomis WF, Smith DW (1990) Molecular phylogeny of Dictyostelium discoideum by protein sequence comparison. Proc Natl Acad Sci USA 87:9093–9097

Maddison DR, Maddison WP (1997) The tree of life project. URL http://phylogeny.arizona.edu/tree/life.html

Naylor GJP, Brown WM (1998) Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. Syst Biol 47:61–76

Nei M (1996) Phylogenetic analysis in molecular evolutionary genetics. Annu Rev Genet 30:371–403

Porter TD (1995) Correlation between codon usage, regional genomic nucleotide composition, and amino acid composition in the cytochrome P-450 gene superfamily. Biochim Biophys Acta 1261:394–400

Russo CA, Takezaki N, Nei M (1996) Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol Biol Evol 13:525–536

Sueoka N (1961) Compositional correlation between deoxyribonucleic acid and protein. Cold Spring Harbor Symp Quant Biol 26:35–43

Swofford DL (1998) PAUP*. Phylogenetic analysis using parsimony (*and other methods), Version 4. Sinauer Associates, Sunderland, MA

Swofford DL, Olson GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: Hillis DM, Moritz G, Mable BK (eds), Molecular systematics, 2nd ed. Sinauer, Sunderland, MA

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680