Reference: *Biol. Bull.* **196:** 345–350. (June 1999)

Genes and Other Samples of DNA Sequence Data for Phylogenetic Inference

MICHAEL P. CUMMINGS[1,*], SARAH P. OTTO[2], AND JOHN WAKELEY[3,†]

[1] *Department of Botany and Plant Sciences, University of California, Riverside, California 92521-1024;* [2] *Department of Zoology, University of British Columbia, 6270 University Blvd. Vancouver, British Columbia V6T 1Z4, Canada; and* [3] *Department of Ecology, Evolution & Natural Resources, Nelson Hall, Rutgers University, P.O. Box 1059, Piscataway, New Jersey 08855-1059*

One of the most basic uses of DNA sequence data in the study of evolution is as a source of information for inferring evolutionary history. Where homology may be difficult to establish, particularly in comparisons of phylogenetic lineages that have diverged relatively early, DNA sequence data may offer distinct advantages over other data, such as morphology. However, DNA sequence data may present its own difficulties, and the sampling properties of DNA sequence data are not well characterized.

To better understand the sampling properties of DNA sequence data in phylogenetic analysis, a series of computational experiments were performed using complete mitochondrial genomes from 10 vertebrate species. These taxa were cow, *Bos taurus;* carp, *Cyprinus carpio;* chicken, *Gallus gallus;* human, *Homo sapiens;* loach, *Crossostoma lacustre;* mouse, *Mus musculus;* rat, *Rattus norvegicus;* harbor seal, *Phoca vitulina;* fin whale, *Balenoptera physalus;* and frog, *Xenopus laevis.* For this study, the mitochondrial genome has some distinct advantages:

1. A number of complete genome sequences are available.
2. A complete genome represents an entire population of sites in a statistical sense.
3. It has a simple history with little or no recombination.
4. It is a convenient size for analysis.
5. It is widely used in systematic studies.

The complete DNA sequences from the mitochondrial genomes of these organisms, exclusive of the control region, were aligned gene by gene and assembled into a data set of 16,075 sites. These data formed the basis of phylogenetic analyses using three methods of phylogenetic tree construction: maximum likelihood (1), parsimony (2), and neighbor-joining (3). These three methods were chosen in part because of their widespread use, but mainly to ascertain whether observed patterns seen were specific to the tree construction methodology or were more general properties of DNA sequence data sampling. Further details of the analyses are given in Cummings *et al.* (4).

As definable and recognizable units, genes represent the most common currency by which DNA sequences are considered; and for reasons related to history, function, and experimental utility, most DNA sequences used in phylogenetic study are, in whole or in part, gene sequences. Therefore, the first question considered was, Do individual genes provide an accurate estimate of whole-genome results?

For these experiments the sequences of all the major genes (those exclusive of tRNA genes) were analyzed by the three phylogenetic methods. Analysis of the complete genomes results in a single common and unambiguously supported tree (Fig. 1), and thus the gene trees could be directly compared to a common tree.

* Current address: The Josephine Bay Paul Center for Comparative Molecular Biology and Evolution, Marine Biological Laboratory, 7 MBL Street, Woods Hole, MA 02543-1015.

† Current address: Department of Organismic and Evolutionary Biology, The Biological Laboratories, 16 Divinity Avenue, Harvard University, Cambridge, MA 02318.

This paper was originally presented at a workshop titled *Evolution: A Molecular Point of View.* The workshop, which was held at the Marine Biological Laboratory, Woods Hole, Massachusetts, from 24–26 October 1997, was sponsored by the Center for Advanced Studies in the Space Life Sciences at MBL and funded by the National Aeronautics and Space Administration under Cooperative Agreement NCC 2-896.

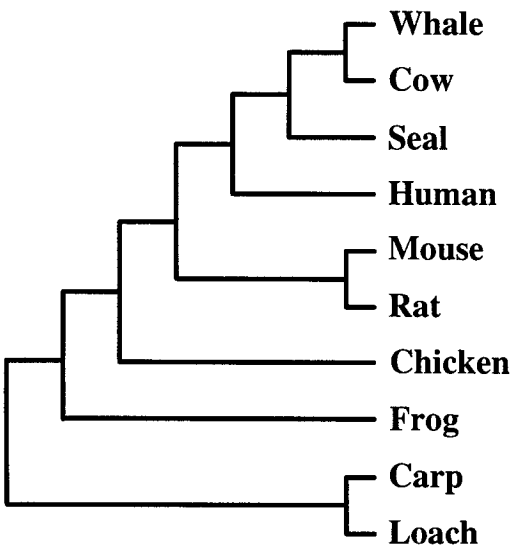

Figure 1. Tree of phylogenetic relationships based on the entire mitochondrial genome, exclusive of the control region, as inferred using maximum likelihood, parsimony, and neighbor-joining.

The major observations from these gene-based experiments were as follows:

1. Few trees inferred from individual genes are identical to the whole-genome tree (3–5 out of 15 genes, depending on the tree construction method; Table I).
2. Twenty different trees were inferred from individual gene sequences across all three construction methods.
3. No single alternative tree was commonly found.
4. No gene sequence gave the same tree for all three methods.
5. Labile branches were not restricted to one part of the tree; *e.g.*, both shallow and deep relationships were unstable (see also references 4 and 5).

There appeared to be some crude relationship between the length of a gene and its chance of leading to a tree identical to the whole-genome tree; indeed, only one gene less than 1111 bp (COIII, 785 bp) gave a tree identical to the whole-genome tree (Table I).

The next question was, If individual gene sequences are not good samples, then how many sites are needed? The answer to this question was obtained simultaneously with the answer to a closely related question, Does it matter how sites are sampled? To answer these questions, two types of random samples were collected and analyzed, and the resulting trees were compared to the whole-genome tree. One sampling scheme is similar to the method of data collection in empirical studies; a site defines one end of a sequence region, and adjacent bases are determined to produce a contiguous sequence (Fig. 2A). This involves the collection of n contiguous sites (where n = 1000, 2000, 3000, . . . , 8000) starting from a random nucleotide position in the genome. The second sampling scheme also involved the collection of n sites, but the sites were individually and independently sampled, without replacement, from random locations throughout the genome (Fig. 2B). By examining multiple collections (1024) of each different sample size, we can determine how many sites are needed to produce a tree identical to that of the whole genome with any chosen level of probability. Simultaneously, by examining the two different sampling schemes, we can determine whether contiguous sites are independently and identically distributed (i.i.d.). Knowing whether contiguous sites are i.i.d. is important, because i.i.d. is a basic assumption of the bootstrap (6), which is a common means of evaluating confidence limits of inferred phylogenetic relationships (7). The reference point in the experiments with these sampling schemes was again the whole-genome tree, but this time each set of samples was evaluated with regard to the proportion of the sample that produced a tree identical to the whole genome tree.

The exact results of these experiments were dependent on the method of tree construction, but several general patterns were evident. The two most fundamental were that many sites are required to have a high probability of producing a tree identical to the whole genome tree, and that samples of contiguous sites do not perform as well as samples of sites

Table I

Results of gene-based sampling experiments

		Phylogenetic Inference Method*		
Gene	Length (bp)	Likelihood	Parsimony	Neighbor-Joining
ATPase8	207			
NADH4L	297			
NADH3	350			
NADH6	561			
ATPase6	687			
COII	705			
COIII	785	×	×	
NADH1	981			
NADH2	1047			
12S rRNA	1111	×	×	
CYTB	1149	×		
NADH4	1387		×	×
COI	1560	×		
16S rRNA	1786	×		×
NADH5	1860			×
Number of distinct topologies		9	12	9

* The symbol × denotes that the topology of the gene tree was identical to that of the genome tree; absence of a symbol means that some other topology was inferred for that gene-method combination (see reference 4 for the alternative gene trees).

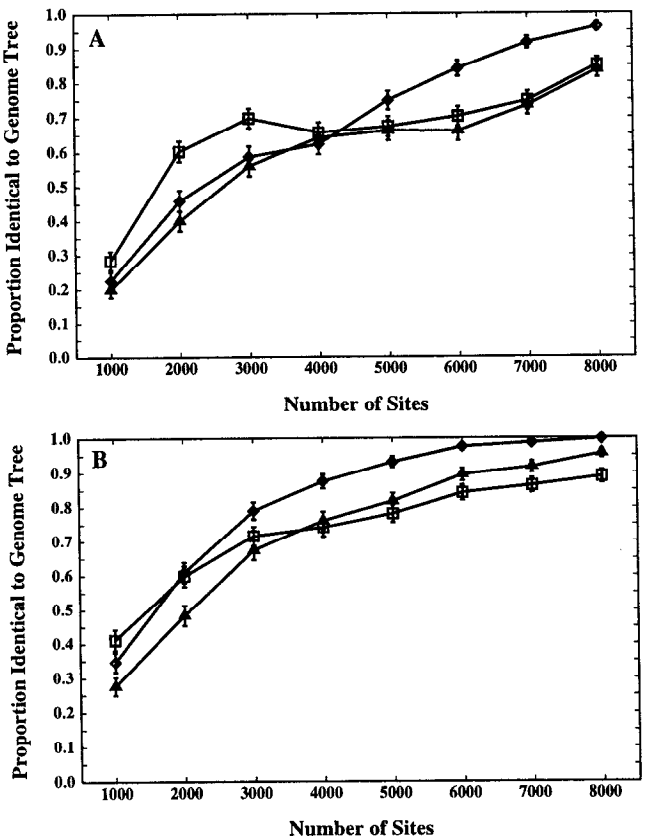

Figure 2. Proportion of trees identical to the whole-genome tree inferred from different sampling schemes. Data points represent mean of 1024 samples of the indicated size, error bars denote 95% confidence intervals for the mean, diamonds represent maximum likelihood, triangles represent parsimony, and squares represent neighbor-joining. (A) Samples of contiguous sites beginning at random locations. (B) Samples of sites individually and independently chosen without replacement from random locations throughout the genome.

that are individually and independently chosen. The observation that the two sampling schemes produce different results is evidence that contiguous sequence data do not meet the i.i.d. assumption.

Taken over the entire study, the principal conclusions are that individual gene sequences are not sufficient samples from which to infer the phylogeny of these taxa; and that contiguous DNA sequence data are not i.i.d. and hence do not meet the basic assumption of the bootstrap. More detail and elaboration of these and other points can be found in references 4 and 5.

Literature Cited

1. **Felsenstein, J. 1981.** Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17:** 368–376.
2. **Fitch, W. M. 1971.** Toward defining the course of evolution: minimal change for a specific tree topology. *Syst. Biol.* **20:** 406–416.
3. **Saitou, N., and M. Nei. 1987.** The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4:** 406–425.
4. **Cummings, M. P., S. P. Otto, and J. Wakeley. 1995.** Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* **12:** 814–822.
5. **Otto, S. P., M. P. Cummings, and J. Wakeley. 1996.** Inferring phylogenies from DNA sequence data: the effects of sampling. Pp. 103–115 in *New Uses for New Phylogenies,* P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith, and S. Nee, eds. Oxford University Press.
6. **Efron, B. 1979.** Bootstrap methods: another look at the jacknife. *Ann. Statist.* **7:** 1–26.
7. **Felsenstein, J. 1985.** Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39:** 783–791.

Discussion

LANDWEBER: Marty Kreitman developed a sequencing method (1) that used a battery of four-cutter restriction enzymes to recognize, on average, 20% of the genome, at best. Your work suggests that method could be resurrected. If one surveyed a 5 kb region with a battery of four-cutter restriction enzymes, that might be better than sequencing 1000 bases of DNA.

CUMMINGS: Kreitman's method of four-cutter analysis ended up being an expensive, laborious method of approximating sequencing. You could only look at a small region at a time using the four-cutter restriction enzymes. When you have to look at a larger region, it does not work out very well. To get at the issue you raise, we considered doing simulated restriction analyses of this whole genome data set. Assuming that there is no problem in establishing homology between fragments, the answer is already in the results presented. The results of a four-cutter approach will, at best, fall between the curve of sequencing everything completely randomly (see Fig. 2B) and the curve for sequencing contiguously (see Fig. 2A). You are not going to do any better than sequencing random sites throughout the whole genome. Any design that is chosen will,

at best, simply place the results between those two curves. For example, you want to know the best design and are only going to sequence 1 kb and have to decide between sequencing a single 1 kb region; 2 × 500 bp regions; or 10 × 100 bp regions. On average, you are better off with 10 × 100 bp regions than with a single contiguous stretch of sequence. I think this gets at the issue.

KATZ: You showed a discouraging table from the individual genes and how they were wrong so often. Did you do any calculation of the branch collapsing? How many steps away is this? We would be a lot more upset if you had fish and humans coming out as sister taxa than if you had changes within mammals.

CUMMINGS: The topological differences covered a pretty broad range, but the implications depend on the metric used. One characteristic of the metric of Robinson and Foulds (2) is that when one taxon is placed on the other end of the tree, relative to the reference tree, the whole tree must be collapsed. Among the notable consistent features in the 20 trees of this study were that the two fish always occurred together as sister taxa, and the two rodents were always together. The relationships among the amniotes and the basic relationships within the mammals showed differences.

SHARP: Do you ever use codons to make these trees?

CUMMINGS: Yes. We split the 11 kb of protein coding sequence from these mitochondrial genomes into first positions, second positions, and third positions of codons. We then analyzed random samples within those three groups separately. It is known, *a priori,* that the second codon position is by far the most conserved position, followed by the first position; and the third position is the most variable (3). One can generate very good trees from just third positions alone, if one uses maximum likelihood. In fact, if you take second positions using parsimony (which are the best for parsimony, because they are the most conserved, having evolved at the slowest rate) and plot the rate of convergence in terms of the portion of trees that are identical to the whole-genome tree based on increasing sample size, and then repeat this with third positions using maximum likelihood, the two curves are quite similar (Fig. 3). The whole notion that third positions are random noise is fallacious. Substitutions proceed in such a way that there is appreciable phylogenetic signal just in third positions alone.

SHARP: That cannot be true.

CUMMINGS: It is true; we have done the experiments (4, 5). Third positions are saturated, in that on average each has undergone one or more substitution events. However, these changes are not random in any true sense. Third positions are only noise, from a practical standpoint, for some methods of phylogenetic analysis. For other methods, such as maximum likelihood, third positions contain a significant amount of information.

KATZ: What is the effect of linkage in broadening your conclusions? Would your expectations change if we were looking at nuclear genes, where there are genes that are not physically linked to one another?

CUMMINGS: Then you get into what Joe (Felsenstein) presented at this meeting (Felsenstein, these proceedings, pp. 343–344)—that different genes can have different histories—and that is why

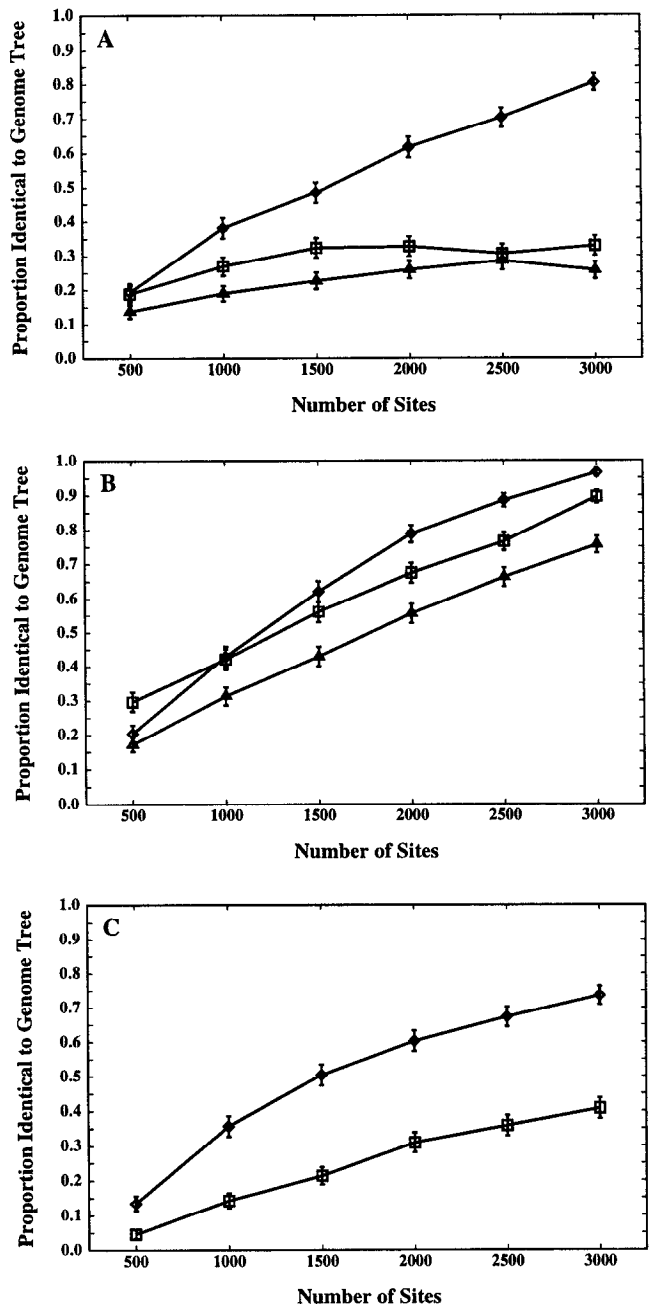

Figure 3. Proportion of trees identical to the whole-genome tree inferred from samples of sites from each codon position class. Data points represent means of 1024 samples of the indicated size, error bars denote 95% confidence intervals for the mean, diamonds represent maximum likelihood, triangles represent parsimony, and squares represent neighbor-joining. (A) First position of codons. (B) Second position of codons. (C) Third position of codons. For third positions analyzed by parsimony, the performance of the branch and bound search approximates an exhaustive search due to the high level of homoplasy in the third position data; consequently this method required an unacceptable length of time to find the shortest tree. Preliminary analyses indicated that for all sample sizes, the number of equally parsimonious trees is generally quite large ($>10^5$), and the resulting proportion of trees identical to the genome tree was very small ($<10^{-5}$).

we wanted to use mitochondrial genomes. An advantage of the mitochondrial genome is that every gene within it has the same history. In this work we are looking at differences in history as inferred from a collection of sites and not at true differences in history. Every site has the same history for the mitochondrial genome.

FORTERRE: Do you assume that the complete tree, which mixes all these wrong trees, can become a good tree?

CUMMINGS: Yes, but there are two ways to look at that. First of all, this is the largest data set ever used in an analysis of these taxa, and it is the best estimate that we have. Secondly, this is the mitochondrial tree, and we are not going to get any more mitochondrial data than the complete genome. One could also say that whether the whole mitochondrial genome tree is really the right tree or not is unimportant. If we do not believe that the whole-genome tree is correct, but we really do think that, for example, the one from cytochrome oxidase subunit two (COII) inferred from neighbor-joining is really the correct tree, what do we do? We disagree with the largest sample that we have available to address the question. If we cast the whole-genome result aside and say that a smaller, seemingly arbitrarily chosen sample is better than the whole genome, I think we are in trouble.

FELSENSTEIN: There is a method that allows the bootstrap to cope with samples that have certain kinds of departure from independent and identically distributed assumption. That is the block bootstrap of Hans Künsch (6). What it does is to say, "Suppose the problem is that neighboring sites (sites within a certain distance of each other) are correlated either in rate or in the particular process." We will then sample a random place and take a block of ten sites starting there, and take a number of samples which is one-tenth as great. Instead of sampling a thousand sites out of a thousand, we will sample a hundred blocks of ten. Have you tried that method, and would it help the bootstrap be a more accurate indication of the variation?

CUMMINGS: No, we have not done experiments along the lines you suggest. This does point out something that is not so dramatic here in the maximum likelihood curves, but can be seen to a greater extent in the neighbor-joining curves. In the curves presented for all analytical methods there is a phenomenon happening at a large scale in these genomes that happened repeatedly in a very consistent manner. In neighbor-joining, for example, at about 3000 sites, it reaches a little peak and then dips. This means that a worse estimate of relationships is obtained on average using 4000 or 5000 sites than with 3000 sites (see Fig. 2A). One must increase the sample of sites to 6000 sites before one gets as good an estimate as with 3000. There is a large-scale phenomenon happening when we take contiguous blocks of 3000, 4000, 5000, 6000 sites. It is unclear, even if a program for your (Felsenstein) suggested method were available, what the appropriate block-size values should be. One might use such a program to explore a range of values to see at what scale variability happens in a particular genome. Based on work of many people (7), all sorts of patterns will occur within codons, for example. There will also be all sorts of interactions on many different levels (8). The scale at which one has to sample sequences to meet the assumptions of the bootstrap is not clear. Furthermore, it is unclear how much the assumption is violated, and whether it makes much difference in terms of bootstrap values and interpretation.

OLSEN: You have asserted that it would be better to take 10 blocks of 100 nucleotides. Did you actually attempt that in simulation?

CUMMINGS: No, but you know that, on average, it is going to be approximately between the results from sampling contiguous sites (Fig. 2A) and from sampling sites individually and independently (Fig. 2B).

OLSEN: I understand that it has to fall between them. The question is, is it a significant step toward independence or is it false security?

CUMMINGS: It depends on how secure you want to be. The only way you can sequence is by taking contiguous blocks. Again, on average you will get better estimates of phylogenetic relationships if you are able to collect samples from different regions of the genome. Although we have not done the experiment you suggest, again, the results will fall between the two curves (Fig. 2).

OLSEN: As a comment, or a reaction, you certainly chose to put a negative spin on your presentation. You tried to recap at the end, but saying that you got twenty different trees out of two million possible is still certainly a small subset. It is particularly depressing in terms of not telling us how far those departures were in terms of either your measure, or symmetric difference, or any of the tree metrics available.

CUMMINGS: All those twenty trees are published (4). One has to choose which particular differences are important, and which are not. Whether the results of a phylogenetic analysis are important depends on your question. I cannot say whether it is a big difference or not; I am not going to make a decision arbitrarily about whether something is important for a particular question. As an aside, I would like to acknowledge that part of the computer code that we used for this work was based, in part, upon code written by Gary (Olsen), Joe (Felsenstein), and their collaborators. My coauthors and I highly modified the code for our purposes. We certainly benefited greatly from their programs.

SHARP: If you had taken the human sequence out of your tree, the rest of it would be unambiguous. We think we are absolutely certain what the relationships between those species would have been if the human were out.

CUMMINGS: That is not true.

SHARP: What would be in question?

CUMMINGS: The relationships that are labile were not restricted to any one part of the tree.

SHARP: That is not my point; I am referring to morphology and everything else. We know the two fish are closer to one another than they are to us. I do not think you are ever going to overturn that. We know that the same is true for the two rodents. You said that you do not care if you get the right tree or not. I am just saying

that if you remove the human sequence, we would know what the right tree was that you were aiming towards.

CUMMINGS: Are you saying that primates were the only disputed group in that tree?

SHARP: The relationship of primates, artiodactyls plus cetaceans and rodents, that would be the only unambiguous one. My question concerns a very similar piece of work by Nei and coworkers who looked at complete mitochondrial sequences and compared their results with an examination of individual genes. I am not sure how your work differs from theirs. I remember that they were far more optimistic in their conclusions and seemed to get the right answer far more often.

CUMMINGS: The paper from Masatoshi Nei's group was an extension of our study in some respects and different in others. Among the differences are restricting examination to protein coding genes; including additional mitochondrial genomes available at the time; and using more tree-reconstruction methods. You can read both papers and interpret them. It depends on what is important to you, whether you think the results are really negative, or if I have put a negative spin on it.

SHARP: My question concerns results, not viewpoints. At least in their paper, they reported what was honestly found. They got the right answer for the majority of genes, and did not just use the three methods that you used.

CUMMINGS: No, their results were not that appreciably different from ours. Are you referring to the paper by Russo *et al.* (9)? Where the two studies were similar in design, they came out with largely the same results.

KUNKEL: Do you have a better model with nuclear genes, or do your results (where you have these twenty different trees) suggest that you need to modify the model?

CUMMINGS: I assume that you are referring to the use of mitochondrial genomes as a model system to address the questions in our study. Would one get a different answer if one were to use complete bacterial genomes, or data with more genes to choose from? If we made experimental design choices, now that there are more vertebrate mitochondrial genome sequences, we would probably get a slightly different answer, although the same general trends would hold. The taxa in this study were convenient; they were the first ten mitochondrial genomes, the only ones available at the time, the only complete non-recombining genomes, and ten was a convenient number to deal with because it provided a lot of variation. A lot of possibilities.

We are all interested in evolutionary history, and do the best we can. The study just shows that how badly one does, or how well one does, depends on how you look at it. For most trees, it does not take very much data to get very close to getting most relationships right. However, it takes a lot of sequence to get it all perfect. Whether that is horrible or great depends on your perspective.

FELSENSTEIN: The alternatives that you gave, of taking a thousand bases or ten chunks of one hundred, aren't very economically realistic. People who sequence (and I've never sequenced anything, I hasten to add; nor will I) tell me that to get four hundred in a chunk instead of one hundred is not much more effort. Can you give us an idea of what the real tradeoff might be? It looks to me, at any rate, that simply counting bases is not going to give you the real choices that you are faced with.

CUMMINGS: I think the overriding concerns, as someone who has sequenced quite a lot, are experimental considerations and convenience. If I were doing the sequencing, I would sequence contiguous chunks. In many cases you are faced with what you are given. I do not think there is a general answer to your question. One cannot say that one is best off sequencing a particular number of blocks each of a particular number of sites. Each situation is going to be different and the experimental design is going to be dictated by practical concerns.

Literature Cited

1. **Kreitman, M., and M. Aguade. 1986.** Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. *Proc. Natl. Acad. Sci. USA* **83:** 3562–3566.
2. **Robinson, D. F., and L. R. Foulds. 1981.** Comparison of phylogenetic trees. *Math. Biosci.* **53:** 131–147.
3. **Nei, M. 1987.** *Molecular Evolutionary Genetics.* Columbia University Press, New York.
4. **Cummings, M. P., S. P. Otto, and J. Wakeley. 1995.** Sampling properties of DNA sequence data in phylogenetic analysis. *Mol. Biol. Evol.* **12:** 814–822.
5. **Otto, S. P., M. P. Cummings, and J. Wakeley. 1996.** Inferring phylogenies from DNA sequence data: the effects of sampling. Pp. 103–115 in *New Uses for New Phylogenies,* P. H. Harvey, A. J. Leigh Brown, J. Maynard Smith, and S. Nee, eds. Oxford University Press.
6. **Künsch, H. R. 1989.** The jackknife and the bootstrap for general stationary observations. *Ann. Stat.* **17:** 1217–1241.
7. **Grantham, R., C. Gautier, M. Gouy, R. Mercier, and A. Pave. 1980.** Codon catalog usage and the genome hypothesis. *Nucleic Acids Res.* **8:** r49–r62.
8. **Peng, C.-K., S. V. Buldyrev, A. L. Goldberger, S. Havlin, F. Sciortino, M. Simons, and H. E. Stanley. 1992.** Long-range correlations in nucleotide sequences. *Nature* **356:** 168–170.
9. **Russo, C. A. M., N. Takezaki, and M. Nei. 1996.** Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. *Mol. Biol. Evol.* **13:** 525–536.