First publ. in: Zoology 108 (2005), 4, pp. 329–336

Magic bullets and golden rules: Data sampling in molecular phylogenetics

Michael P. Cummings[a,*], Axel Meyer[b,*]

[a]Center for Bioinformatics and Computational Biology, University of Maryland, College Park, MD 20742, USA
[b]Lehrstuhl für Zoologie und Evolutionsbiologie, Department of Biology, University of Konstanz, D-78457 Konstanz, Germany

Received 13 July 2005; received in revised form 22 September 2005; accepted 23 September 2005

Abstract

Data collection for molecular phylogenetic studies is based on samples of both genes and taxa. In an ideal world, with no limitations to resources, as many genes could be sampled as deemed necessary to address phylogenetic problems. Given limited resources in the real world, inadequate (in terms of choice of genes or number of genes) sequences or restricted taxon sampling can adversely affect the reliability or information gained in phylogenetics. Recent empirical and simulation-based studies of data sampling in molecular phylogenetics have reached differing conclusions on how to deal with these problems. Some advocated sampling more genes, others more taxa. There is certainly no 'magic bullet' that will fit all phylogenetic problems, and no specific 'golden rules' have been deduced, other than that single genes may not always contain sufficient phylogenetic information. However, several general conclusions and suggestions can be made. One suggestion is that the determination of a multiple, but moderate number (e.g., 6–10) of gene sequences might take precedence over sequencing a larger set of genes and thereby permit the sampling of more taxa for a phylogenetic study.

Keywords: Phylogenetics; Data sampling; Gene sampling; Taxon sampling

Introduction

In the ideal world, a molecular systematist having unlimited resources and time could collect DNA sequences for as many genes from as many taxa as deemed necessary to answer a particular phylogenetic problem. In the real world, a sample of one or only a few DNA sequences is assumed to be representative of the entire genome for a subset of taxa. These taxa in turn are assumed to constitute the relevant taxonomic context for making accurate inferences regarding evolutionary relationships. For several reasons, these two assumptions may be incorrect and, hence, can lead to inaccurate phylogenetic inference. Although there is universal agreement that more data are better, with limited resources compromises have to be made. For phylogenetic problems, this compromise is cast in terms of a tradeoff between sampling in the two dimensions of the phylogenetic data matrix: length of sequences and number of taxa (Graybeal, 1998; Mitchell et al., 2000). Here we review the results of recent studies that have examined the effects of data sampling in phylogenetics including examination of complete genomes.

Although appropriate choices of taxa and gene regions can improve the results of phylogenetic analyses,

*Corresponding authors.
E-mail addresses: mike@umiacs.umd.edu (M.P. Cummings), axel.meyer@uni-konstanz.de (A. Meyer).

all experimental choices that an investigator makes regarding sampling in a molecular phylogenetic study are constrained: (i) by the previous sampling of the evolutionary process itself; and (ii) by the choice of genetic markers that have been used before for a particular taxon. With regard to the first point this means: sampling manifested through the processes of speciation, which will tend to lead to short branches and the processes of extinction, which will tend to lead to longer branches. Both processes affect the number of lineages that are available for a particular phylogenetic problem. Among the consequences of the macroevolutionary events are that some long branches will always remain long (e.g., those leading to lungfish, coelacanths, Lissamphibia or Microsporidia), and some short branches are indeed short regardless of which sampling choices are made (e.g., species that are part of adaptive radiations such as the extremely closely related species of cichlid fishes from Lake Victoria). Additional processes, such as mutation, natural selection (e.g., codon usage bias), drift, gene duplication and loss, and gene conversion further expand or contract the variation and "gene space" that is available for phylogenetic problems. Together these processes create the specific universe of taxa and gene and genome sequences from which investigators can choose samples to use as the basis for evolutionary inference.

The effects of sampling on phylogenetic inference have been studied using both empirical and simulated data, both of which have strengths and limitations. Empirical data are the result of myriad biological processes some of which are poorly understood which makes it difficult to extrapolate from individual studies to potential future ones. Given the historical aspect of evolution, empirical data can be very idiosyncratic which further makes extrapolation difficult. These properties result in empirical data having a mixture of characteristics that provide increased realism to experimental designs. In contrast, simulated data are the result of much simplified models designed to emulate particular characteristics of biological processes. Hence, simulated data are often moderately general. The control of specific characteristics with simulated data can be advantageous in experimental designs. Studies based on empirical data and studies based on simulated data thus can provide complementary perspectives on analytical problems.

Gene sampling

That every gene within a genome will have experienced the same species history is a common a priori assumption. But clearly sampling only a usually miniscule portion of the genome can produce, at least potentially, a large variance in results. Moreover, hybridization, gene duplications, polyploidization, horizontal gene transfer, and other evolutionary events in recent history can lead to differences in gene histories. Additionally, particularly ancient symbiotic and other past reticulation events further highlight the possibility that gene histories can and do differ within the same organism. It is becoming increasingly obvious that the "tree of life" might resemble more a "net of life" or "ring of life" (Rivera and Lake, 2004); therefore, especially for phylogenetic problems that involve the relationships among different kingdoms of life, particular care must be taken in selecting genes for phylogenetic studies. For such phylogenetic problems that involve vast evolutionary distances it appears that the determination of entire genomes is the best option in the attempt to study evolutionary relationships.

The whole-genome option is typically not available to most evolutionary biologists interested in some more restricted phylogenetic problems, such as relationships among salamanders or flowering plants. Most often, a single or a small set of genes needs to be selected at the onset of a study. This choice of gene(s) is more often than not predominantly influenced by history (e.g., past experimental studies) and perceived technical considerations rather than a priori knowledge regarding the sufficiency of information for enabling phylogenetic inference for the problem at hand. There are favorite genes (e.g., 16SrRNA, 18SrRNA, 28SrRNA, COI, *cytb*, *rbc*L) that have long been used for data in phylogenetic studies, and some (e.g., c-mos, EF1-a, *rag1*) that have become popular much more recently (Fig. 1). These favorites grew in popularity often because "universal" PCR-primers were developed for them early during the PCR-revolution of molecular systematics (e.g., Kocher et al., 1989), and the combination of the resulting lowered technical hurdles and the sociology of science contributed to the predominant role that a relatively small set of genes now plays in the field of molecular systematics.

How to best analyze several orthologous genes for a given set of taxa has been debated – the issue is whether genes should get analyzed individually, and consensus trees should be constructed, or whether the gene sequences should be concatenated and analyzed in aggregate (de Queiroz et al., 1995). In combined analyses several approaches can be used, but two dominate: (i) to apply different models of nucleotide substitution or other parameters to the individual genes (or other partitions) making up the combined data, or (ii) to use model averaging. Although no universal agreement has been reached on the question of how to best analyze several genes for one taxon, it seems fair to say that most molecular systematists now favor the concatenated approach (reviewed in Gadagkar et al.,

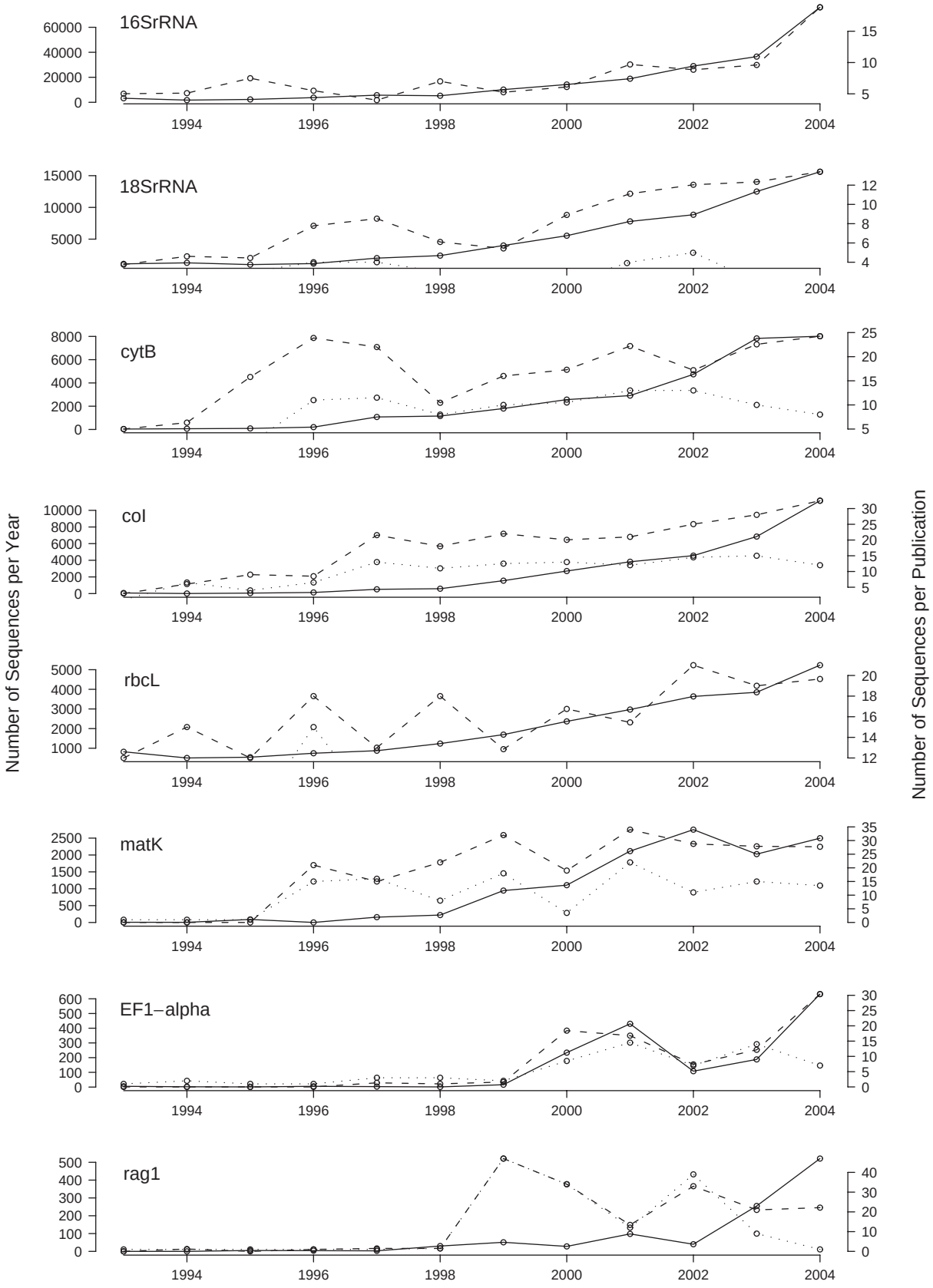

Fig. 1. Plots depicting the total number of sequences per year (solid line; left axis), and the mean (dashed line) and median (dotted line) number of sequences per publication per year (right axis), for genes commonly used in systematic studies from 1993 to 2004. The number of sequences per publication per year provides an estimate of the number of taxa per publication per year for studies using the gene, because the results are presented for each gene separately. GenBank records were collected for each gene, filtered to include only those with whole or substantial partial gene sequences, and then parsed to generate the numbers used in the calculations.

2005), because the problems associated with building consensus trees from a set of individual trees seem more severe because information is lost and some, possibly strongly supported, clades might not be found in the consensus. Recent simulation studies (Gadagkar et al., 2005) and analyses based on empirical data (Driskell et al., 2004) find that a concatenated data set of 6–10 genes will result in trees that are largely accurate. An interesting outcome of one study (Gadagkar et al., 2005) is that, at least for a simulated phylogenetic problem based on the mammalian radiation, it mattered little which genes were chosen even without attempts to use multiple models. This result is similar to that found in a study of complete genomes from several yeast species by Rokas et al. (2003).

Rokas and Carroll (2005) specifically assessed the relative contribution of gene sampling and taxon sampling using genome-based data for 14 yeast species. Their results show that accuracy of inferred relationships is significantly improved with increased gene numbers, but there was no significant effect with increased taxon numbers, in contrast to what has been suggested before (Hillis, 1996).

The amount of DNA sequence information necessary to resolve a particular phylogenetic issue depends on the problem at hand. Hence, no "golden rule" can be derived from previous studies with the possible exception that the information within a single gene is insufficient for robust inference of phylogenetic relationships. This general insufficiency of single genes has sometimes been well documented, for example, the gene coding for cytochrome B (Graybeal, 1993; Meyer, 1994). Not surprisingly more sequence data per taxon, often including sequences of multiple genes, have been demonstrated to substantially improve accuracy and support in phylogenetic inference (Cao et al., 1994; Cummings et al., 1995, 1999; Otto et al., 1996; Nei et al., 1998; Mitchell et al., 2000; Poe and Swofford, 1999; Rokas and Carroll, 2005; Rokas et al., 2003; Yoder and Irwin, 1999; Zardoya and Meyer, 1996). There were notable early efforts to comparatively evaluate the phylogenetic utility of a number of different genes (Friedlander et al., 1992; Graybeal, 1994; Russo et al., 1996). Subsequently, the decreasing costs of collecting data, the increasing standards and recognition of the power of more data have led to a general trend toward more data per taxon including multiple genes in phylogenetic studies. This trend is manifested by the transition from single gene to multi-gene to genome-based studies that recently accelerated (Cummings et al., 1995; Zardoya and Meyer, 1996; Adachi et al., 2000; Pollock et al., 2000; Babteste et al., 2002; Matsuoka et al., 2002; Rokas et al., 2003; Daubin et al., 2003; Lerat et al., 2003; Philippe et al., 2004; Vogl et al., 2003; Chen et al., 2004; Goremykin, 2004; Battistuzzi et al., 2004).

Heterogeneity of sequence data

Evolutionary processes make a DNA sequence heterogeneous in various ways and at various scales (Karlin and Brendel, 1993). Accounting for some of this heterogeneity has been the basis for the increasing parameter richness and complexity of models for DNA sequence evolution that are used in phylogenetic inference. Important types of heterogeneity influencing phylogenetic inference include among-site substitution rate variation (Wakeley, 1996; Yang, 1996). Extreme substitution variation among sites is very much evident for most genes or other genomic regions (e.g., variation among different codon positions, different structural regions of rRNA, or sites in the mitochondrial genome displacement loop). Therefore, each gene or other sequence region sampled from a genome provides a sample of nucleotide sites that have evolved at a variety of different rates and, hence, will have different levels of information regarding phylogenetic relationships, as has been described for rRNA genes (Woese, 1987; Van de Peer et al., 2000).

Although within-gene rate variation is extreme, there is clear evidence for the non-independence in the evolution of DNA (e.g., Cummings et al., 1995, 1999; Otto et al., 1996; Comeron and Kreitman, 1998; Smith and Hurst, 1999; Kelchner, 2002), which has (at least philosophical) consequences for inferring relationships (under the assumption of character independence for each site within codons, within protein domains and within secondary structure elements of rRNA genes) and the support of those relationships (Cummings et al., 1995, 1999; Otto et al., 1996; Huelsenbeck and Nielsen, 1999; Galtier, 2004). The correlation in substitution rates between adjacent sites can be either positive or negative. An example of negative correlation has been found in protein-coding sequences where, as a consequence of the genetic code and purifying selection, infrequently substituted second positions in codons are followed by more frequently substituted third positions. The evolutionary non-independence among sites decreases the effective amount of information available in contiguous sequences, which, in turn, may reduce the accuracy of inferred relationships. Furthermore, this non-independence might falsely increase apparent support for relationships (e.g., bootstrap values), even in cases where those relationships are incorrect. Thus, using sequences from genes that are dispersed throughout the genome (where local selection or base-composition environments might differ) has beneficial consequences in terms of improving the accuracy of inferred relationships and support for those relationships (Cummings et al., 1995, 1999; Otto et al., 1996; Mitchell et al., 2000; Koepfli and Wayne, 2003; Galtier, 2004).

As mentioned above, the genome-scale option will typically not be available, and even complete genomes

are, of course, not going to entirely eliminate all potential problems in determining phylogenetic relationships; on the contrary, they might even exacerbate them (Phillips et al., 2004). Included among these problems are those related to lineage-specific biases in either base composition or rate of evolution. Different phylogenetic methods are known to succumb to these effects more easily or to be more robust to these non-historical biases. One (old) solution to this problem is to code nucleotides as purines and pyrimidines (RY-coding) or other ways such as focusing on the most slowly evolving sites (Brinkmann and Philippe, 1999) to counter the negative consequences of relative compositional variability (RCV) (Phillips and Penny, 2003). Other attempts (e.g., higher weighting of transversions compared to typically more frequent transitions) have long been used and can be incorporated into many of the commonly used phylogenetic methods by using more complex models (e.g. LogDet or Hadamard transformed distances) and other means to counteract base compositional biases. These methods of data coding, weighting and phylogenetic models all strive to increase (give more weight to) the strongest, presumably most reliable and historically correct, signal. Yet, we are not aware of tests that will be able to tell when this has been achieved.

Questions involving recent speciation events or rapid radiations might require a large data set of quickly evolving genomic regions (or alternative approaches such as microsatellite, SNPs, or AFLPs which sample a larger portion of the genome) to resolve evolutionary relationships.

Taxon sampling

The issue of taxon sampling in the context of data sampling in phylogenetics will likely not be answerable with a 'golden rule', because it is very much problem dependent. It has been known for some time that the levels of homoplasy (e.g., as measured by a consistency index) will increase with increasing numbers of taxa (Sanderson and Donoghue, 1989). Thus suggestions that larger sets of taxa will result in more accurate/robust phylogenies came as a surprise (Hillis, 1996) and it remains debated how generally this effect is expected to work.

The most studied and controversial questions in data sampling for phylogenetics are which taxa should be represented, how many taxa are included, and how they should be distributed phylogenetically. The consequences, in terms of accuracy and support for inferred phylogenetic relationships, attributable to the taxonomic sample depend on the details of the biological situation as has been shown in empirical studies involving lemurs (Yoder and Irwin, 1999), land plants (Soltis et al., 1999; Rydin and Kallersjo, 2002), birds (Omland et al., 1999; Saunders and Edwards, 2000; Johnson, 2001; Braun and Kimball, 2002), noctuoid moths (Mitchell et al., 2000), xenarthrans (anteaters, armadillos and sloths) (Delsuc et al., 2002), heterokont algae (Goertzen and Theriot, 2003), and many simulations (Sanderson and Donoghue, 1989; Kim, 1996; Graybeal, 1998; Hillis, 1998; Rannala et al., 1998; Poe and Swofford, 1999; Rosenberg and Kumar, 2001, 2003; Pollock et al., 2002; Zwickel and Hillis, 2002; Hillis et al., 2003; Poe, 2003). Some of the results of these studies are apparently contradictory, because of factors relating to different performance measures, interactions with specific inference methods, and the nature of the taxa added or deleted. In some cases, addition of taxa can be helpful in resolving phylogenetic relationships. For example, the long accepted view that long-branch subdivision through the addition of taxa is one way to increase accuracy of phylogenetic relationships (Hendy and Penny, 1989) is supported by many studies (e.g., Brinkmann et al., 2004). However, there are exceptions that depend on where the added taxa intersect long branches (Poe, 2003; Poe and Swofford, 1999). The use of model-based phylogenetic analysis methods, with appropriately fitted models, reduces the long-branch attraction problem to some extent and simultaneously increases the value of taxon addition for long-branch subdivision. Furthermore, the addition of taxa that are internal to monophyletic groups generally increases support (e.g. bootstrap values) for such groups. However, addition of taxa can have the opposite effect if the models are incorrect (Poe, 2003).

Although rarely considered in molecular phylogenetics, sampling multiple individuals for a species has been advocated as being necessary for testing the paraphyly or polyphyly of species (Funk and Omland, 2003). This intraspecific sampling is crucial in studies of closely related species, as ~23% of 2319 species surveyed exhibit paraphyly or polyphyly (Funk and Omland, 2003).

Increased taxon sampling has clear, demonstrably beneficial consequences for model parameter estimation and tests associated with phylogenetic analyses. This has been demonstrated for determining the root node of a large clade (Sanderson, 1996), the power of the relative-rate test (Robinson, 1998), the estimation of substitution rate parameters (Sullivan et al., 1999; Pollock and Bruno, 2000), and the estimation of ancestral character states (Salisbury and Kim, 2001).

Completeness of datasets

Experimental constraints can lead to missing data, which result in a sparse data matrix. Sparse matrices are

particularly characteristic of data sets assembled by mining sequence databases. The original sequence in each database entry represents data collected for a variety of reasons and, hence, all of the data available for subsequent synthetic phylogenetic analysis rarely fills a full rectangular matrix. Studies examining the use of sparse matrices to infer phylogenetic relationships (Wiens, 1998, 2003a, b; Driskell et al., 2004) have shown that, in general, adding taxa with missing data within monophyletic groups was less likely to decrease accuracy but adding taxa with missing data in situations involving long-branch attraction was more likely to decrease accuracy. Elimination of whole rows and/or columns, which is sometimes done to reduce sparseness of a matrix, can result in substantial loss of information and precludes establishing relationships for the eliminated taxa. The problems sometimes manifested with sparse matrices do not result from the proportion of missing data per se, but rather from the insufficiency of the data available for phylogenetic inference: a subtle, but important distinction. It has been demonstrated that as much as 95% of the data can be missing for some taxa, but if the remaining 5% is sufficiently informative, relationships can be accurately inferred (Wiens, 2003b). Therefore, it is best to use all of the available data, use appropriate model-based analytical methods, and assess the support for phylogenetic inferences.

Conclusions

The practitioner of molecular systematics will want to know which genes to sequence and whether sufficient sequence data and taxa have been sampled to accurately address particular phylogenetic problems. We cannot offer a "magic bullet", because for every problem there is a level of sampling sufficient for accurate and robust results. The choice of genes should be guided by which have been shown to be informative for phylogenetic problems of similar 'depth'. Unless severe problems are discovered, analysis of combined (concatenated) data, particularly when using multiple models or model averaging, would be expected to have the highest success in solving phylogenetic problems.

Choices regarding data in molecular phylogenetics in many regards follow the general patterns of sampling in other domains. In keeping with the law of large numbers, increasing sample size has demonstrably beneficial consequences in most cases. As matrices increase in size to fulfill the goal of increasing taxon representation, sparseness will also increase, particularly with the use of constructing data sets through the mining of sequence databases. With these come the additional challenges of larger scale phylogenetic analysis (Sanderson and Driskell, 2003). With the increasing use of phylogenetic relationships to provide the context for much of biology, consideration of the choices of data sampling and their consequences can help make for informed decisions and interpretations.

Acknowledgements

We thank Adam Bazinet and Maile Neel for comments and assistance with the figure, and anonymous reviewers for suggestions. MPC is supported by grants from the US National Science Foundation, Department of Energy, and National Aeronautics and Space Administration. AM is supported by grants from the Deutsche Forschungsgemeinschaft, the EU, the HFSP, and the University of Konstanz.

References

Adachi, J., Waddell, P.J., Martin, W., Hasegawa, M., 2000. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J. Mol. Evol. 50, 348–358.

Babteste, E., Brinkmann, H., Lee, J.A., Moore, D.V., Sensen, C.W., Gordon, P., Duruflé, L., Gaasterland, T., Lopez, P., Müller, M., Philippe, H., 2002. The analysis of 100 genes supports the grouping of three highly divergent amoebae: *Dictyostelium*, *Entamoeba*, and *Mastigamoeba*. Proc. Natl. Acad. Sci. USA 99, 1414–1419.

Battistuzzi, F.U., Feijao, A., Hedges, S.B., 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evol. Biol. 4, 44.

Braun, E.L., Kimball, R.T., 2002. Examining basal avian divergences with mitochondrial sequences: model complexity, taxon sampling, and sequence length. Syst. Biol. 51, 614–625.

Brinkmann, H., Philippe, H., 1999. Archea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol. Biol. Evol. 16, 817–825.

Brinkmann, H., Denk, A., Zitzler, J., Joss, J.J., Meyer, A., 2004. Complete mitochondrial genome sequences of the South-American and the Australian lungfishes: testing of the phylogenetic performance of mitochondrial data sets for phylogenetic problems in tetrapod relationships. J. Mol. Evol. 59, 834–838.

Cao, Y., Adachi, J., Janke, A., Paabo, S., Hasegawa, M., 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J. Mol. Evol. 39, 519–527.

Chen, W.-J., Orti, G., Meyer, A., 2004. Novel evolutionary relationship among four fish model systems. Trends Genet. 20, 424–438.

Comeron, J.M., Kreitman, M., 1998. The correlation between synonymous and nonsynonymous substitutions in Drosophila: mutation, selection or relaxed constraints? Genetics 150, 767–775.

Cummings, M.P., Otto, S.P., Wakeley, J., 1995. Sampling properties of DNA sequence data in phylogenetic analysis. Mol. Biol. Evol. 12, 814–822.

Cummings, M.P., Otto, S.P., Wakeley, J., 1999. Genes and other samples of DNA sequence data for phylogenetic inference. Biol. Bull. 196, 345–350.

Daubin, V., Moran, N.A., Ochman, H., 2003. Phylogenetics and the cohesion of bacterial genomes. Science 301, 829–832.

Delsuc, F., Scally, M., Madsen, O., Stanhope, M.J., de Jong, W.W., Catzeflis, F.M., Springer, M.S., Douzery, E.J., 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Mol. Biol. Evol. 19, 1656–1671.

de Queiroz, A., Donoghue, M.J., Kim, J., 1995. Separate versus combined analysis of phylogenetic evidence. Annu. Rev. Ecol. Syst. 26, 657–681.

Driskell, A.C., Ane, C., Burleigh, J.G., McMahon, M.M., O'Meara, B.C., Sanderson, M.J., 2004. Prospects for building tree of life from large sequence databases. Science 306, 1172–1174.

Friedlander, T.P., Regier, J.C., Mitter, C., 1992. Nuclear gene sequences for higher level phylogenetic analysis: 14 promising candidates. Syst. Biol. 41, 483–490.

Funk, D.J., Omland, K.E., 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34, 397–423.

Gadagkar, S.R., Rosenberg, M.S., Kumar, S., 2005. Inferring species phylogenies from multiple genes: concatenated sequence tree versus consensus gene tree. J. Exp. Zool. Part B: Mol. Dev. Evol. 304B, 64–74.

Galtier, N., 2004. Sampling properties of the bootstrap support in molecular phylogeny: influence of non-independence among sites. Syst. Biol. 53, 38–46.

Goertzen, L.R., Theriot, E.C., 2003. Effect of taxon sampling, character weighting, and combined data on the interpretation of relationships among the heterokont algae. J. Phycol. 39, 423–443.

Goremykin, V.V., 2004. The chloroplast genome of *Nymphaea alba*: whole-genome analyses and the problem of identifying the most basal angiosperm. Mol. Biol. Evol. 21, 1445–1454.

Graybeal, A., 1993. The phylogenetic utility of cytochrome b: lessons from bufonid frogs. Mol. Phylogenet. Evol. 2, 256–269.

Graybeal, A., 1994. Evaluating the phylogenetic utility of genes: a search for genes informative about deep divergences among vertebrates. Syst. Biol. 43, 174–193.

Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Syst. Biol. 47, 9–17.

Hendy, M.D., Penny, D., 1989. A framework for the quantitative study of evolutionary trees. Syst. Zool. 38, 297–309.

Hillis, D.M., 1996. Inferring complex phylogenies. Nature 383, 130–131.

Hillis, D.M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47, 3–8.

Hillis, D.M., Pollock, D.D., McGuire, J.A., Zwickl, D.J., 2003. Is sparse taxon sampling a problem for phylogenetic inference? Syst. Biol. 52, 124–126.

Huelsenbeck, J.P., Nielsen, R., 1999. Effect of nonindependent substitution on phylogenetic accuracy. Syst. Biol. 48, 317–328.

Johnson, K.P., 2001. Taxon sampling and the phylogenetic position of Passeriformes: evidence from 916 avian cytochrome b sequences. Syst. Biol. 50, 128–136.

Karlin, S., Brendel, V., 1993. Patchiness and correlations in DNA sequences. Science 259, 677–680.

Kelchner, S.A., 2002. Group II introns as phylogenetic tools: structure, function, and evolutionary constraints. Am. J. Bot. 89, 1651–1669.

Kim, J., 1996. General inconsistency conditions for maximum parsimony: effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45, 363–374.

Kocher, T.D., Thomas, W.K., Meyer, A., Edwards, S.V., Paabo, S., Villablanca, F.X., Wilson, A.C., 1989. Dynamics of mitochondrial evolution in animals: amplification and sequencing with conserved primers. Proc. Natl. Acad. Sci. USA 86, 6196–6200.

Koepfli, K.P., Wayne, R.K., 2003. Type I STS markers are more informative than cytochrome b in phylogenetic reconstruction of the Mustelidae (Mammalia: Carnivora). Syst. Biol. 52, 571–593.

Lerat, E., Daubin, V., Moran, N.A., 2003. From gene trees to organismal phylogeny in prokaryotes, a case for gamma-proteobacteria. PLoS Biol. 1, 101–109.

Matsuoka, Y., Yamazaki, Y., Ogihara, Y., Tsunewaki, K., 2002. Whole chloroplast genome comparison of rice, maize, and wheat: implications for chloroplast gene diversification and phylogeny of cereals. Mol. Biol. Evol. 19, 2084–2091.

Meyer, A., 1994. Shortcomings of the cytochrome b gene as a molecular marker. Trends Ecol. Evol. 9, 278–280.

Mitchell, A., Mitter, C., Regier, J.C., 2000. More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepioptera). Syst. Biol. 49, 202–224.

Nei, M., Kumar, S., Takahashi, K., 1998. The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc. Natl. Acad. Sci. USA 95, 12390–12397.

Omland, K.E., Lanyon, S.M., Fritz, S.J., 1999. A molecular phylogeny of the new world orioles (Icterus): the importance of dense taxon sampling. Mol. Phylogenet. Evol. 12, 224–239.

Otto, S.P., Cummings, M.P., Wakeley, J., 1996. Inferring phylogenies from DNA sequence data: the effects of sampling. In: Harvey, P.H., Brown, A.J.L., Smith, J.M., Nee, S. (Eds.), New Uses for New Phylogenies. Oxford University Press, Oxford, pp. 103–115.

Philippe, H., Snell, E.A., Bapteste, E., Lopez, P., Holland, P.W.H., Casane, D., 2004. Phylogenomics of eukaryotes: the impact of missing data on large alignments. Mol. Biol. Evol. 21, 1740–1752.

Phillips, M.J., Penny, D., 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogenet. Evol. 28, 171–185.

Phillips, M.J., Delsuc, F., Penny, D., 2004. Genome-scale phylogeny and the detection of systematic biases. Mol. Evol. Biol. 21, 1455–1458.

Poe, S., 2003. Evaluation of the strategy of long-branch subdivision to improve the accuracy of phylogenetic methods. Syst. Biol. 52, 423–428.

Poe, S., Swofford, D.L., 1999. Taxon sampling revisited. Nature 389, 299–300.

Pollock, D.D., Bruno, W.J., 2000. Assessing an unknown evolutionary process: effect of increasing site-specific knowledge through taxon addition. Mol. Biol. Evol. 17, 1854–1858.

Pollock, D.D., Eisen, J.A., Doggett, N.A., Cummings, M.P., 2000. A case for evolutionary genomics and the comprehensive examination of sequence biodiversity. Mol. Biol. Evol. 17, 1776–1788.

Pollock, D.D., Zwickl, D.J., McGuire, J.A., Hillis, D.M., 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51, 664–671.

Rannala, B., Huelsenbeck, J.P., Yang, Z., Nielsen, R., 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47, 702–710.

Rivera, M.C., Lake, J.A., 2004. The ring of life provides evidence for a genome fusion origin of eukaryotes. Nature 431, 134–137.

Robinson, M., 1998. Sensitivity of the relative-rate test to taxonomic sampling. Mol. Biol. Evol. 15, 1091–1098.

Rokas, A., Carroll, S.B., 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22, 1337–1344.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.

Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98, 10751–10756.

Rosenberg, M.S., Kumar, S., 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52, 119–124.

Russo, C.A., 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.

Rydin, C., Kallersjo, M., 2002. Taxon sampling and seed plant phylogeny. Cladistics 18, 485–513.

Salisbury, B.A., Kim, J.H., 2001. Ancestral state estimation and taxon sampling density. Syst. Biol. 50, 557–564.

Sanderson, M.J., 1996. How many taxa must be sampled to identify the root node of a large clade? Syst. Biol. 45, 168–173.

Sanderson, M.J., Donoghue, M.J., 1989. Patterns of variation in levels of homoplasy. Evolution 43, 1781–1795.

Sanderson, M.J., Driskell, A.C., 2003. The challenge of constructing large phylogenetic trees. Trends Plant Sci. 8, 374–379.

Saunders, M.A., Edwards, S.V., 2000. Dynamics and phylogenetic implications of mtDNA control region sequences in new world jays (Aves: Corvidae). J. Mol. Evol. 51, 97–109.

Smith, N.G.C., Hurst, L.D., 1999. The effect of tandem substitutions on the correlation between synonymous and nonsynonymous rates in rodents. Genetics 153, 1395–1402.

Soltis, P.S., Soltis, D.E., Wolf, P.G., Nickrent, D.L., Chaw, S.M., Chapman, R.L., 1999. The phylogeny of land plants inferred from 18S rDNA sequences: pushing the limits of rDNA signal? Mol. Biol. Evol. 16, 1774–1784.

Sullivan, J., Swofford, D.L., Naylor, G.J.P., 1999. The effect of taxon sampling on estimating rate heterogeneity parameters of maximum-likelihood models. Mol. Biol. Evol. 16, 1347–1356.

Van de Peer, Y., Baldauf, S.L., Doolittle, W.F., Meyer, A., 2000. An updated and comprehensive rRNA phylogeny of (crown) eukaryotes based on rate-calibrated evolutionary distances. J. Mol. Evol. 51, 565–576.

Vogl, C., Badger, J., Kearney, P., Li, M., Clegg, M., Jiang, T., 2003. Probabilistic analysis indicates discordant gene trees in chloroplast evolution. J. Mol. Evol. 56, 330–340.

Wakeley, J., 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. Trends Ecol. Evol. 11, 158–163.

Wiens, J.J., 1998. Does adding characters with missing data increase or decrease phylogenetic accuracy? Syst. Biol. 47, 625–640.

Wiens, J.J., 2003a. Incomplete taxa, incomplete characters, and phylogenetic accuracy: is there a missing data problem? J. Vertebr. Paleontol. 23, 297–310.

Wiens, J.J., 2003b. Missing data, incomplete taxa, and phylogenetic accuracy. Syst. Biol. 52, 528–538.

Woese, C.R., 1987. Bacterial evolution. Microbiol. Rev. 51, 221–271.

Yang, Z., 1996. Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. 11, 367–372.

Yoder, A.D., Irwin, J.A., 1999. Phylogeny of the Lemuridae: effects of taxon and character sampling on resolution of species relationships within Eulemur. Cladistics 15, 351–361.

Zardoya, R., Meyer, A., 1996. The phylogenetic performance of mitochondrial protein coding genes in resolving relationships among vertebrates. Mol. Biol. Evol. 13, 933–942.

Zwickel, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–598.