Published in final edited form as:

Mol Phylogenet Evol. 2007 October ; 45(1): 81–88. doi:10.1016/j.ympev.2007.04.022.

The Effect of Branch Lengths on Phylogeny: an Empirical Study Using Highly Conserved Orthologs from Mammalian Genomes

Austin L. Hughes¹ and Robert Friedman²

¹Department of Biological Sciences, University of South Carolina, Columbia SC 29208 USA

²Bioinformatics Center, University of Connecticut, Storrs CT 06066

Abstract

Phylogenetic analyses were applied to 269 families of putative orthologs represented by a single member in the genomes of human, mouse, dog, and chicken. Five methods were used: maximum parsimony (NP), neighbor-joining (NJ) with Poisson and Gamma distances; and maximum likelihood (ML) with JTT and JTT + gamma models. When applied to the concatenated sequence of all families, all methods strongly supported a tree in which mouse branched before human and dog. In analyses of individual families, the same topology was supported more than any other. Although there was evidence of an increased rate of amino acid replacement in the mouse lineage in comparison to the other two mammals, there was no evidence that support for the mouse's basal position was due to long-branch attraction; rather, this topology was seen in the families with the lowest rate variation among the three mammalian branches. In families with highly divergent mouse sequences, ML with both JTT and JTT + gamma and NJ with the gamma distance tended to support a topology in which the dog, rather than the mouse, branched first. Thus, in these data, a tendency of long and short branches to cluster together ("opposite-branch attraction") seemed to be more of a problem than long-branch attraction.

Keywords

long-branch attraction; mammalian phylogeny; opposite-branch attraction; phylogenetic methods

Molecular sequence data have opened the possibility of addressing phylogenetic questions with increased power, given the extensive new information such data contain regarding the evolutionary history of organisms (Nei and Kumar 2000; Felsenstein 2003). A complete genome sequence, in particular, provides the maximal amount of data on a given species (Blair et al. 2005). However, given that, at least in the foreseeable future, the number of completely sequenced genomes is likely to remain limited, systematic biologists are often faced with a trade-off. If the goal is to include in a phylogenetic analysis data on as many taxa as possible, the amount of sequence data for each taxon is likely to be limited. On the other hand, if the goal is to maximize the amount of data for each taxon included and thus to rely on complete or nearly complete genome sequences, the number of taxa available for analysis is likely to be limited.

Address correspondence to: Austin L. Hughes, Ph.D., Department of Biological Sciences, University of South Carolina, Coker Life Sciences Bldg., 700 Sumter St. Columbia SC 29208 USA, Tel: 1-803-777-9186, Fax: 1-803-777-4002, austin@biol.sc.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

The question of whether or not a phylogenetic analysis using a limited number of taxa is likely to be erroneous (the problem of “taxon sampling”) has been much debated (Hillis 1998; Poe 1998; Rannala et al. 1998; Mitchell et al. 2000; Rosenberg and Kumar 2001; Zwickl and Hillis 2002; Pollock et al. 2002; Hillis et al. 2003; Rosenberg and Kumar 2003; DeBry 2005). An argument in favor of increased taxon sampling is that additional taxa serve to break up long branches and thereby help prevent the recovery of erroneous topologies through the phenomenon known as long-branch attraction (LBA), whereby certain phylogenetic methods tend to cluster long branches together (Felsenstein 1978; Hendy and Penny 1989; Hillis 1998). LBA is generally held to be particularly problematic for the maximum parsimony (MP) method, whereas maximum likelihood (ML) is much less prone to LBA (Anderson and Swofford 2004). It has also been proposed that the neighbor-joining (NJ) method can be prone to LBA, at least under certain conditions (Fares, Byrne, and Wolfe 2006). On the other hand, there is evidence that the ML method can be prone to a phenomenon of “long-branch repulsion,” which perhaps might better be termed “opposite-branch attraction” (OBA), whereby unusually long branches are attracted to unusually short branches (Yang 1996; Pol and Siddall 2001).

The relationships among the orders of eutherian (placental) mammals represents an unresolved phylogenetic question to which issues of both taxon sampling and LBA are relevant. Early molecular analyses tended to support topologies in which Rodentia branched outside the other major placental orders, including Primates and Carnivora (Figure 1A) (Easteal 1990; Li et al. 1990). On the other hand, in more recent analyses using a larger number of taxa, Rodentia and Primates tended to cluster together (Figure 1B) (Murphy et al. 2001). Some other recent analyses with an intermediate number of taxa have placed Rodentia outside the other placental orders (Misawa and Janke 2003; Misawa and Nei 2003; Cannarozzi et al. 2007; but see Jørgensen et al. 2005).

One possible factor in these contradictory results may be an accelerated rate of molecular evolution in rodents in comparison to other mammalian orders, particularly primates, as shown by a variety of methods (Gu and Li 1992). Thus, it might be argued that the tendency for rodents to cluster outside the other placental orders is due to an inability of certain methods of phylogenetic analysis to accommodate substantial rate variation among branches in a phylogeny and/or to LBA. If so, it might be predicted that those phylogenetic methods that are sensitive to rate variation among branches or are prone to LBA will be more likely to cluster the rodents outside the other placental orders. However, it is important to heed the warning of Anderson and Swofford (2004) and not merely to assume that LBA is responsible whenever a phylogenetic analysis yields an unexpected or undesired result.

Here we use protein sequences encoded by putatively orthologous genes from three complete mammalian genomes in order to test for associations between the topology of the phylogenetic tree, the method of phylogenetic analysis, and the variation in branch length. The three mammalian species analyzed were human (order Primates), dog (order Carnivora), and mouse (order Rodentia); and sequences from a bird genome, that of the chicken, were used as an outgroup to root the tree. Note that the chicken provides a more appropriate outgroup for the purposes of the present study than would a closer outgroup, such as a marsupial mammal. Because our analyses focused on rate differences among the three mammalian taxa, a more distant outgroup is preferable because it allows time for sufficient numbers of differences to have accumulated in order to distinguish true rate differences from stochastic error.

There are three possible phylogenies for the rooted tree (Figure 1). After identifying gene families having a single representative in each of the four genomes, we analyzed each family separately, as well as the concatenated sequence, using three phylogenetic reconstruction methods: maximum parsimony (MP), neighbor joining (NJ), and maximum

likelihood (ML). Our purpose was not to determine definitively which of these phylogenies is correct. Rather, using a limited number of taxa, we sought to examine empirically the properties of the sequence data sets that favored a given topology using a given phylogenetic method.

A problem with identifying putative orthologs from each of a set of genomes (“panorthologs”) merely by the presence of a single representative in each of the genomes analyzed is that a certain number of genes that are paralogous rather than orthologous (“pseudo-orthologs”) may be included. Pseudo-orthology can arise naturally as a result of independent deletion events in different lineages (Figure 2). On the other hand, pseudo-orthology can also occur artifactually if true orthologs are omitted due to errors in genome assembly or annotation. Likewise, errors in identifying gene family members by homology search may potentially lead to misidentification of orthologs; such errors can occur because homology searches do not always accurately reflect phylogenetic relationships (Koski and Golding 2001). Finally, the problem of pseudo-orthologs is likely to be particularly acute in cases where a complete genome is not available for each species in the analysis, since without genomic information it may often be difficult to distinguish paralogs from orthologs. In order to eliminate pseudo-orthologs, one needs to know in advance the true organismal phylogeny. But, in cases of phylogenetic analysis, one does not know this phylogeny; therefore, there is no obvious *a priori* criterion for eliminating pseudo-orthologs. The presence of pseudo-orthologs thus represents an additional problem which, along with LBA/OBA, can potentially mislead phylogenetic analyses based on putative orthologs from a set of complete genomes.

Methods

Sequence data

The genome sequence data for four species of vertebrates was obtained from the Ensembl web site (<http://www.ensembl.org>). Each of the species is associated with a genome assembly (and version number) as follows: human *Homo sapiens* (30.35c); mouse *Mus musculus* (30.33f); dog *Canis familiaris* (30.1b); and chicken *Gallus gallus* (30.1f). Each dataset was filtered so that only unique loci (genes) remained. This filtering limited the data set to one transcript (chosen at random) per physical locus and also removed unmapped genes. In order to ensure that corresponding transcripts were chosen for each species, we examined all alternative transcripts manually at selected loci.

Gene families

The BLASTCLUST software (version 2.2.9; Altschul et al. 1997) was used to cluster genes into families. This program both identifies sequence homology and then assembles sequences into families by the single-linkage method. We applied BLASTCLUST to predicted amino acid sequences using the default settings, except for the following: E (expect value) = 10^{-6} and minimum sequence match of 20% across a minimum of 40% of the amino acid sequence alignment. These relatively relaxed values allowed us to identify the more distant *Gallus* homologs, while still choosing relatively conserved proteins (Hughes et al. 2005).

In order to eliminate pseudo-orthologs wrongly identified as orthologs by the single-link method used by BLASTCLUST, we used the protein sequence homology search program BLASTP (with $E = 10^{-30}$) to search our complete database of protein translations for additional homologs of each member of all families having one member in each of the four genomes. When additional homologs were obtained, we conducted a phylogenetic analysis (see below) of the extended family including these additional homologs. When this analysis supported the hypothesis that the genes in the original four-member family were not true orthologs, that family was excluded from further analysis; seven families were excluded as a result of this procedure.

For purposes of analysis, we retained 269 protein families having a single representative in each of the four genomes. The sequences are provided as online supplementary material; the Ensembl accession numbers are provided in Supplementary Table S1. Members of each protein family were aligned using ClustalX (version 1.83; Thompson et al., 1997). In all phylogenetic analyses, any site at which the alignment postulated a gap in any sequence was excluded from the data matrix. Because of the latter procedure, we excluded from phylogenetic analyses any alternatively spliced exon that might happen to occur in the transcript chosen for one of the species but not in all of the others.

Phylogenetic analyses

This subset of the data was also concatenated into a single dataset for phylogenetic analysis. After aligned sites were removed with any gaps present, the concatenated subset contained 115,221 aligned amino acid sites.

Phylogenetic trees were constructed for each family separately and for the concatenated amino acid sequence. Phylogenetic analysis was applied at the amino acid level only because synonymous nucleotide sites were saturated with changes in comparisons between mammals and chicken. Three methods of phylogenetic analysis were used: maximum parsimony (MP); neighbor joining (NJ); and maximum likelihood (ML). For MP the branch and bound algorithm was used. NJ trees were based on both the Poisson-corrected amino acid distance and the gamma-corrected amino acid distance, which takes into account rate variation among sites (Nei and Kumar 2000). The MEGA3 program was used for MP and NJ analyses (Kumar, Tamura, and Nei 2004). ML trees were constructed using the Phylip software package (*proml* program; Felsenstein 1989) with a JTT model of sequence evolution (Jones, Taylor, and Thornton 1992). ML analyses were conducted assuming a constant rate of evolution among sites and assuming that rates varied among sites following a gamma distribution. The TREEPUZZLE program (Strimmer and van Haeseler 1996) was used to estimate the shape parameter (α) of the gamma distribution for both NJ and ML analyses. Reliability of clustering patterns in the trees based on concatenated sequences was tested by bootstrapping (Felsenstein 1985); 1000 bootstrap samples were used.

Other statistical analyses

In order to assess the association between rate variation among branches and the results of phylogenetic analyses, we computed for each family the proportion of amino acid difference (p) between each of the three mammalian taxa and the outgroup (chicken). Then we computed the coefficient of variation (CV) among these three p values (where CV is defined as the ratio of the standard deviation to the mean). CV provides a unitless measure of variation, corrected for differences among families with respect to the amount of accumulated amino acid differences. The uncorrected proportion of amino acid differences (p) was used, since amino acid distances corrected using a model of amino acid sequence evolution are highly correlated with p but always have a greater stochastic error of estimation. For example, in the present data, the correlations between p and the Poisson-corrected distances were all greater than 0.98, while those between p and the gamma-corrected distances were all greater than 0.93. However, the variances of both Poisson-corrected and gamma-corrected distances were substantially greater than that of p . In preliminary analyses, we conducted analyses using the Poisson-corrected and gamma-corrected amino acid distances instead of p ; the results were essentially identical (data not shown).

Because the variable analyzed (CV of p and standardized residuals from regression) were not normally distributed, we used the nonparametric Kruskal-Wallis test to test for differences among group medians (Hollander and Wolfe 1973). In preliminary analyses, parametric

analysis of variance applied to means yielded essentially identical results, reflecting the well-known robustness of one-way analysis of variance to violation of the assumption of normality.

Results

Phylogenetic Analyses

The phylogeny of concatenated amino acid sequences was reconstructed by the following methods: MP; NJ based on the Poisson-corrected amino acid distance; NJ based on the gamma-corrected amino acid distance; ML assuming a constant rate; and ML assuming that rates vary according to a gamma distribution. The shape parameter (α) for the gamma distribution was estimated to be 0.58. All methods yielded the same topology (topology A; Figure 1A; and in every case bootstrap support for the internal branch was 100%. Thus, in the concatenated data, all methods supported the phylogenetic relationships proposed by Li et al. (1990), rather than those proposed by Murphy et al. (2001). As a consequence, support for the former topology could not be attributed to an artifact arising from any particular method.

When the results of analyses for the 269 individual families were examined, the number of families supporting topology A (Figure 1A) was higher than that supporting topology B (Figure 1B) or topology C (Figure 1C) for all methods (Table 1). The NJ method with the Poisson distance yielded support for topology A in the highest number of families (177), with MP supporting the next highest number (162) (Table 1). In every case except ML with rate variation (JTT + gamma), the number of families supporting topology A was a majority of all families, whereas in the case of ML with JTT + gamma, a plurality of families supported topology A (Table 1). The five methods agreed on topology in 122 (45.4%) of the 269 families. Of these 122 families, in 82 (67.2%) families, topology A was favored; in 22 (18.0%) families, topology B was favored; and in 18 (14.8%) families, topology C was favored.

Variation among Branches

For the 269 proteins, the mean ($\pm$ S.E.) proportion of amino acid difference (p) between human and chicken was 0.302 ± 0.011 (median 0.280); the mean p between dog and chicken was 0.309 ± 0.010 (median 0.285); and the mean p between mouse and chicken was 0.309 ± 0.011 (median 0.281). Among the p values for all comparisons of the mammals with chicken, 74% were less than 0.400; and 92% were less than 0.600.

When CV of p between each mammal and chicken was compared among families for which different topologies were supported by the different methods, a significant difference in median CV was found for all methods (Table 2). For every method, the families in which topology A was preferred had the lowest median CV, the families in which topology C was preferred had the highest median CV (Table 2). Similarly, considering only the 122 families on which all methods agreed, the lowest median CV was observed in the families for which topology A was supported, while the highest median CV was observed in families for which topology C was supported (Table 2). Thus, all methods showed a tendency to support topology A in the families showing the least variation in evolutionary rate among branches.

Median CV with respect to p differed significantly among families categorized with respect to the outcome of the two constant-rate methods (Table 3). Consistent with the results shown in Table 2, the set of families for which both constant-rate methods favored topology A showed the lowest median CV except for a small number (5) of families in which ML supported topology B and NJ supported topology C (Table 3). We similarly categorized families based on the results of the NJ and ML methods assuming a gamma distribution of rates. Here again, there was a significant difference in median CV among categories, with families for which

both methods favored topology A showing the lowest median CV except for a small number (3) of families in which ML supported topology B and NJ supported topology C (Table 3).

Mouse vs. Dog Sequence Divergence

In order to examine which taxa contributed most to rate variation, we compared p between each of the taxa and chicken. In 175 (65.1%) families, p between mouse and chicken was greater than that between dog and chicken; in 21 (7.8%) families, p between mouse and chicken was equal to that between dog and chicken; and in 73 (27.1%) families, p between mouse and chicken was less than that between dog and chicken. The preponderance of cases in which p between mouse and chicken exceeded that between dog and chicken was highly significant ($P < 0.001$; sign test). Similarly, p between mouse and chicken was greater than that between human and chicken in 180 (66.9%) families; p between mouse and chicken was equal to that between human and chicken in 19 (6.7%) families; and p between mouse and chicken was less than that between human and chicken in 70 (26.0%) families ($P < 0.001$; sign test). By contrast, p between dog and chicken was greater than that between human and chicken in 130 (48.3%) families; p between dog and chicken was equal to that between human and chicken in 30 (11.2%) families; and p between dog and chicken was less than that between human and chicken in 109 (40.5%) families (n.s., sign test). These results supported the hypothesis of an overall acceleration in the rate of amino acid replacement mouse compared to the other two mammal species.

Figures 3A-B show a plot of p between mouse and chicken vs. p between dog and chicken. Many points fell on or close to a straight line, and an R^2 of 0.890 ($P < 0.001$) indicated an overall strong linear relationship. The regression equation was $Y = 0.016 + 0.949X$. However, certain families were clearly outliers to this trend (Figure 3). Many of the outliers were included in the following categories: families for which both NJ with the Poisson model and ML with JTT supported topology B (Figure 3A); families for which NJ with the Poisson model supported topology A, but ML with JTT supported topology B (Figure 3A); families for which both NJ with the gamma model and ML with JTT + gamma supported topology B (Figure 3B); and families for which NJ with the gamma model supported topology A, but ML with JTT + gamma supported topology B (Figure 3B).

We used the standardized residual (SRES) from the regression of p between mouse and chicken on p between dog and chicken (Figure 3) as a measure of deviation from a linear relationship between these quantities. NJ with the gamma model and ML with both JTT and JTT + gamma models showed a significant difference among the three tree topologies with respect to median SRES (Table 2). With all of these methods, by far the highest median SRES was found in families for which the phylogenetic analysis supported topology B (Table 2). These results imply that these methods were likely to choose topology B when the mouse sequence was unusually divergent relative to the dog sequence. Note that this pattern is the opposite of that which would be caused by LBA. If LBA were occurring, one would expect topology A to be favored when the mouse was unusually divergent.

When we categorized families based on the results of NJ and ML assuming a constant rate across sites, there was a significant difference among categories with respect to median SRES (Table 3). The highest median SRES was observed in families for which NJ favored topology A, but ML favored topology B (Table 3). This category of families, including many with high SRES, is illustrated in Figure 3A. Similarly, when we categorized families based on the results of the NJ and ML methods assuming a gamma distribution of rates, there was a highly significant difference among categories with respect to median SRES (Table 3). The highest median SRES was seen in the families for which both methods favored topology B; and the next highest median SRES was seen in the families for which NJ favored topology A, but ML favored topology B (Table 3).

Discussion

Phylogenetic analyses of concatenated amino acid sequences from 269 gene families having a single representative in the genomes of human, dog, mouse, and chicken supported a phylogeny according to which mouse branched outside the other mammals. This topology, here designated topology A (Figure 1A) was that supported by Li et al. (1990), in contrast to topology B (Figure 1B), supported by Murphy et al. (2001). By using a number of different methods of phylogenetic analysis and by examining rate variation among branches, we tested hypotheses that have been proposed to account for Topology A; namely that topology A results from biases of certain phylogenetic methods and/or from long-branch attraction (LBA).

In our analyses, topology A was supported by the maximum parsimony (MP) method; by neighbor-joining (NJ) using the Poisson distance; by NJ using the gamma distance; by the maximum likelihood (ML) method assuming a constant rate of evolution across sites; and by the ML method assuming that rate variation among sites followed a gamma distribution. Thus, support for this topology could not be attributed to an artifact of one particular method or model of sequence evolution. Likewise, when the same methods of phylogenetic analysis were applied to individual families, more families supported topology A than the other two possible topologies with every method.

We examined characteristics of the data sets for each family in order to identify the factors causing the phylogenetic analyses to uncover different topologies in individual families. One factor that might influence topology is rate variation among branches of the tree. As a measure of the extent of rate variation among the three mammalian taxa, we computed the coefficient of variation (CV) of the proportion of amino acid difference (p) between each mammalian taxon and the outgroup (chicken). The results showed that, for all phylogenetic methods, median CV was consistently lowest in the families for which topology A was supported. Similarly, the standard residual of the regression of p between mouse and chicken vs. p between dog and chicken was consistently low when topology A was supported by all methods except MP. These results thus contradicted the hypothesis that support for topology A was in general due to the inability to accommodate rate variation among branches, such as would result from an enhanced rate of amino acid replacement in the rodent lineage.

Unusual patterns of sequence divergence were seen in families in which both ML and NJ supported topology B or in which ML supported topology B while NJ supported topology A. Assuming a constant rate of evolution across sites, families in which ML supported topology B but NJ supported topology A showed unusually high rates of amino acid sequence evolution in mouse compared to dog. Similarly, assuming that rate variation among sites followed the gamma distribution, families in which both methods supported topology B or ML supported topology B but NJ supported topology A showed unusually high rates of evolution in the mouse. These results suggest that opposite-branch attraction (OBA) was much more likely to occur in these families than was long-branch attraction (LBA). If LBA were occurring, the families with highly divergent mouse sequences would be likely to support topology A, at least with MP and NJ methods. In fact, the opposite pattern was seen, especially when NJ was used. In NJ as in ML, highly divergent mouse sequences were associated with topology B.

That OBA would occur in the case of ML might be expected from the known tendencies of this method (Nei and Kumar 2000). However, the results showed evidence that OBA was occurring even in the NJ method, when the gamma-corrected distance was used. This interpretation was supported by the high median value of the standard residual of the regression of p between mouse and chicken vs. p between dog and chicken, in families where both NJ and ML supported topology B (Table 3). Our results thus suggest that OBA may be a more serious problem in phylogenetic studies than is LBA.

Unusually long branches in these families might result from differences in evolutionary rates, but in some cases they might result from the use of pseudo-orthologs. As mentioned previously (see Methods), we attempted to eliminate such cases arising artifactually as a result of homology search; but not all cases arising from differential deletion of genes (Figure 2) or errors in gene prediction and/or genome assembly would be eliminated by the approach we used. Nonetheless, our phylogenetic analysis of the concatenated amino acid sequences of all 269 families suggested that the inclusion of a certain proportion of pseudo-orthologs or otherwise anomalous families is not problematic if data from a large number of genes are used. The dominant signal seen in the majority of families evidently was sufficiently strong to overcome any contrary signals.

Because only three mammalian species were involved, the present analyses cannot resolve the controversial question of the branching order of the major lineages of placental mammals (Murphy et al. 2001; Misawa and Janke 2003; Cannarozzi et al. 2007). Nonetheless, our results help to clarify the issues in the debate, by showing that the phylogeny obtained by Li et al. (1990) (Tree A; Figure 1) cannot plausibly be attributed to either the biases of certain methods of phylogenetic analysis and/or to LBA. Rather, our results suggest that OBA may have been a factor in analyses that have obtained the alternative phylogeny (Tree B; Figure 1).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This research was supported by grant GM43940 from the National Institutes of Health to A.L.H.

Literature Cited

- Altschul SF, Madden TT, Schäffer AA, Zhang J, Zhang A, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 1997;25:3389–3402. [PubMed: 9254694]
- Anderson FE, Swofford DL. Should we be worried about long-branch attraction in real data sets? : Investigations using metazoan 18S rRNA. *Mol Phyl Evol* 2004;33:440–451.
- Blair JE, Shah P, Hedges SB. Evolutionary sequence analysis of complete eukaryotic genomes. *BMC Bioinformatics* 2005;6:53. [PubMed: 15762985]
- Cannarozzi G, Schneider A, Gonnet G. A phylogenomic study of human, dog, and mouse. *PLoS Comput Biol* 2007;3(1):e2.10.1371/journal.pcbi.0030002 [PubMed: 17206860]
- Debry RW. The systematic component of phylogenetic error as a function of taxonomic sampling under parsimony. *Syst Biol* 2005;54:432–440. [PubMed: 16012109]
- Easteal S. The pattern of mammalian evolution and the relative rate of molecular evolution. *Genetics* 1990;124:165–173. [PubMed: 2307352]
- Fares NA, Byrne KP, Wolfe KH. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of *Saccharomyces* species. *Mol Biol Evol* 2006;23:245–253. [PubMed: 16207937]
- Felsenstein J. Cases in which parsimony and compatibility methods will be positively misleading. *Syst Zool* 1978;27:401–410.
- Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 1985;39:783–791.
- Felsenstein J. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 1989;5:164–166.
- Felsenstein, J. *Inferring Phylogenies*. Sinauer Associates; Sunderland MA: 2003.
- Gu X, Li W-H. Higher rates of amino acid substitution in rodents than in humans. *Mol Phyl Evol* 1992;1:211–214.

- Hendy MD, Penny D. A framework for the quantitative study of evolutionary trees. *Syst Zool* 1989;38:297–309.
- Hillis DM. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst Biol* 1998;47:3–8. [PubMed: 12064238]
- Hillis DM, Pollock DD, McGuire JA, Zwickl DJ. Is sparse taxon sampling a problem for phylogenetic inference? *Syst Biol* 2003;52:124–126. [PubMed: 12554446]
- Hollander, M.; Wolfe, DA. *Nonparametric Statistical Methods*. Wiley; New York: 1973.
- Hughes AL, Ekollu V, Friedman R, Rose JR. Gene family content-based phylogeny of prokaryotes: the effect of criteria for inferring homology. *Syst Biol* 2005;54:268–276. [PubMed: 16012097]
- Jones DT, Taylor WR, Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci* 1992;8:275–282. [PubMed: 1633570]
- Jørgensen FG, Hobolth A, Hornshøj H, Bendixen C, Fredholm M, Schierup MH. Comparative analysis of protein coding sequences from human, mouse, and the domesticated pig. *BMC Biology* 2005;3:2.10.1186/1741-7007-3-2 [PubMed: 15679890]
- Koski LB, Golding GB. The closest BLAST hit is often not the nearest neighbor. *J Mol Evol* 2001;52:540–542. [PubMed: 11443357]
- Kumar S, Tamura K, Nei M. MEGA3: integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. *Brief Bioinform* 2004;5:150–163. [PubMed: 15260895]
- Li W-H, Gouy M, Sharp PM, O'hUigin C, Yang Y-W. Molecular phylogeny of Rodentia, Lagomorpha, Primates, Artiodactyla, and Carnivora and molecular clocks. *Proc Natl Acad Sci USA* 1990;87:6703–6707. [PubMed: 2395871]
- Misawa K, Janke A. Revisiting the Glires concept – phylogenetic analysis of nuclear sequences. *Mol Phyl Evol* 2003;28:320–327.
- Misawa K, Nei M. Reanalysis of Murphy et al.'s data gives various mammalian phylogenies and suggests overcredibility of Bayesian trees. *J Mol Evol* 2003;57:S290–S296. [PubMed: 15008427]
- Mitchell A, Mitter C, Regier GC. More taxa or more characters revisited: combining data from nuclear protein-encoding genes for phylogenetic analyses of Noctuoidea (Insecta: Lepidoptera). *Syst Biol* 2000;49:202–224. [PubMed: 12118405]
- Murphy WJ, Eizirik E, O'Brien SJ, Madsen O, Scally M, Doady CJ, Teeling E, Ryder OA, Stanhope MJ, de Jong WW, Springer MS. Resolution of the early placental mammal radiation using Bayesian phylogenetics. *Science* 2001;294:2348–2351. [PubMed: 11743200]
- Nei, M.; Kumar, S. *Molecular Evolution and Phylogenetics*. Oxford University Press; New York: 2000.
- Poe S. Sensitivity of phylogeny estimation to taxonomic sampling. *Syst Biol* 1998;47:18–31. [PubMed: 12064237]
- Pol D, Siddall ME. Biases in maximum likelihood and parsimony: a simulation approach to a 10-taxon case. *Cladistics* 2001;17:266–281.
- Pollock DD, Zwickl DJ, McGuire JA, Hillis DM. Increased taxon sampling is advantageous for phylogenetic inference. *Syst Biol* 2002;51:664–671. [PubMed: 12228008]
- Rannala B, Huelsenbeck JP, Yang Z, Nielsen R. Taxon sampling and the accuracy of large phylogenies. *Syst Biol* 1998;47:702–710. [PubMed: 12066312]
- Rosenberg MS, Kumar S. Incomplete taxon sampling is not a problem for phylogenetic inference. *Proc Natl Acad Sci USA* 2001;98:10751–10756. [PubMed: 11526218]
- Rosenberg MS, Kumar S. Taxon sampling, bioinformatics, and phylogenomics. *Syst Biol* 2003;52:119–124. [PubMed: 12554445]
- Strimmer K, von Haeseler A. Quartet puzzling: a quartet maximum-likelihood method for reconstructing tree topologies. *Mol Biol Evol* 1996;13:964–969.
- Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Research* 1997;25:4876–4882. [PubMed: 9396791]
- Yang Z. Phylogenetic analysis using parsimony and likelihood methods. *J Mol Evol* 1996;42:294–307. [PubMed: 8919881]
- Zwickl DJ, Hillis DM. Increased taxon sampling greatly reduces phylogenetic error. *Syst Biol* 2002;51:588–598. [PubMed: 12228001]

Figure 1. Possible phylogenetic hypotheses for the relationships of human, mouse, and dog, rooted with chicken.

Figure 2. Hypothetical scenario illustrating differential deletions (indicated by “X”) that might lead to a family with one member in each of four genomes (chicken, human, mouse, and dog) which are not true orthologs (“pseudo-orthologs”).

Figure 3.

Plots of the proportion of amino acid difference (p) between mouse and chicken against that between dog and chicken in 269 gene families. Families for which the ML method supported topology B (Figure 1) while NJ supported topology A are indicated by red crosses, while those for which both ML and NJ supported topology B are indicated by green crosses. (A) Methods assuming a constant rate across sites (Poisson distance in NJ and JTT model in ML); (B) Methods assuming that rate variation among sites follows a gamma distribution (gamma-corrected distance in NJ, JTT + gamma in ML).

Table 1
Numbers of families supporting different topologies in phylogenetic analyses by different methods.

Topology ¹	Method			
	MP	NJ	ML	
		Poisson	Gamma	JTT + gamma
A	162 (65.1%) ²	177 (65.8%)	150 (55.8%)	149 (55.4%)
B	51 (20.5%)	60 (22.3%)	77 (28.6)	73 (27.1%)
C	36 (14.4%)	32 (11.9%)	42 (15.6%)	47 (17.5%)
none	20	--	--	--

¹ See Figure 1.

² Percentages (in parentheses) represent the percentage of resolved trees supporting a given topology.

Table 2
Median values of CV (coefficient of variation in p for mammal-chicken comparisons) and SRES (standard residual from regression of p mouse-chicken vs. p dog-chicken) for trees supported by different methods.

Method	Tree ^J			P (Kruskal-Wallis)
	A	B	C	
CV	MP	0.040	0.044	0.074
	NJ			0.001
	Poisson	0.038	0.055	0.097
	Gamma	0.039	0.043	0.074
	ML			0.001
	JTT	0.035	0.046	0.072
SRES	JTT+gamma	0.034	0.045	0.066
	All (consensus)	0.034	0.046	0.113
	MP	0.165	0.152	0.138
	NJ			n.s.
	Poisson	0.132	0.152	0.062
	Gamma	0.072	0.260	0.020
	ML			0.001
	JTT	0.070	0.254	0.109
	JTT+gamma	0.067	0.356	0.119
	All (consensus)	0.069	0.270	0.119

^JFigure 1.

Table 3
Median values of CV (coefficient of variation in p for mammal-chicken comparisons) and SRES (standard residual from regression of p mouse-chicken vs. p dog-chicken) for trees categorized by support of ML and NJ methods.

Model	Tree supported ¹	N	CV	SRES
Constant	ML A	NJ A	133	0.034
		NJ B	12	0.043
		NJ C	4	0.539
	ML B	NJ A	30	0.050
		NJ B	38	0.046
		NJ C	5	0.030
	ML C	NJ A	14	0.047
		NJ B	10	0.106
		NJ C	23	0.108
	P (Kruskal-Wallis)			< 0.001
Gamma	ML A	NJ A	99	0.034
		NJ B	16	0.035
		NJ C	5	0.054
	ML B	NJ A	31	0.049
		NJ B	46	0.044
		NJ C	3	0.009
	ML C	NJ A	20	0.053
		NJ B	15	0.048
		NJ C	34	0.081
	P (Kruskal-Wallis)			0.004