Journal of Phylogenetics & Evolutionary Biology

Som, J Phylogen Evolution Biol 2013, 1:3
DOI: 10.4172/2329-9002.1000116

Research Article

Open Access

Genome-Scale Approach and the Performance of Phylogenetic Methods

Anup Som*

Center of Bioinformatics, Institute of Interdisciplinary Studies, University of Allahabad, Allahabad-211002, India

Abstract

The use of genome-scale approach in phylogenetic analysis is imperative in order to resolve evolutionary relationships over large taxon sets and deep phylogenetic divergences. But yet it is not clear what are the strengths and weaknesses of the various phylogenetic methods or which one should be preferred under genome-scale approach. In this article, the performance of five major phylogenetic methods is evaluated under genome-scale approach using biologically realistic simulated data. The following phylogenetic methods are considered; Bayesian, maximum likelihood (ML), neighbor joining (NJ), NJ maximum composite likelihood (NJ-MCL), and maximum parsimony (MP). Simulation results show that probabilistic methods (i.e., Bayesian and ML methods) are much more accurate than the NJ-MCL, MP and NJ methods. Concerning the consistency of methods, ML is consistent than other methods. This analysis shows that the NJ-MCL, MP, and NJ methods are fast (i.e., computationally efficient), but their accuracy and consistency are very poor compared to Bayesian and ML methods. On the other hand, the Bayesian method is an accurate one, but less consistent than the ML method, and it takes much longer execution time. Therefore, based on the accuracy, consistency and computational efficiency the ML method is the preferred algorithm under genome-scale approach. In addition to the methods performance, this study has investigated several important aspects of genome-scale phylogeny; such as how concatenations of longest and smallest genes make effect on the method's performance, how much datasets are needed to recover the true tree (i.e. true evolutionary history of a group of species or genes), and whether more genes or more characters are important. These are explained in the result section.

Keywords: Genome-scale approach; Phylogenetic methods; Computer simulation; Topological distance; Accuracy; Consistency

Introduction

Reconstruction of evolutionary history from multiple genes is routinely conducted using genome-scale approach where individual gene sequences are concatenated head-to-tail to form a super gene alignment. Improved accuracy of phylogenetic inference through the concatenation of multiple sequences from the same taxon is expected on theoretical grounds [1,2] and has been found in many studies [3-9]. The use of genome-scale approach in phylogenetic analysis is widely applied in order to resolve evolutionary relationships over large taxon sets and deep phylogenetic divergences with greater resolution [10,11]. It has been proposed that a well-resolved Tree of life can be achieved through concatenation of genes [12]. Judging by recent phylogenetic analyses using concatenated genes, the tendency is to combine data by default, in the hope that weight of corroborative evidence will resolve any kind of conflicts [4-6,10]. However, multigene datasets suffer from systematic errors such as within-site-rate variation [13] and long-branch attraction artifacts [14], and statistical methods for extracting information from such data remains limited [15-18]. In this case, analysis of individual partitions (phylogenies are inferred separately for each data set and a consensus tree determined from these separate trees), in addition to combined analysis, is also necessary [19].

A number of studies have been conducted to investigate what the strengths and weaknesses of each method are, or which should be preferred in given situation [20-25]. But, despite of the extensive use of the genome-scale approach, studies comparing the performance of phylogenetic methods under a genome-scale approach are lacking. Advances in both computer and algorithm speed have allowed us to simulate and analyze several thousand data sets, and provided a thorough look at the performance of the various methods.

The study reported here has several purposes: first, to evalthe performance of Bayesian, ML, NJ-MCL, MP, and NJ methods under the genome-scale approach and to find out the most accurate method; second, to examine the effect of addition of a single gene to an existing concatenation on the methods performance; third, to investigate how many datasets are needed to recover the true tree; fourth, to check whether the number of genes or the nsumber of characters is more important.

Here performance of the methods is measured based on three criteria; accuracy[1], consistency[2] and computational efficiency [26,27], and the problem is studied using biologically realistic simulated DNA datasets.

Materials and Methods

Phylogeny reconstruction methods

The performance of five phylogenetic tree reconstruction methods was examined in a genome-scale approach; Bayesian, ML, NJ-MCL, MP, and NJ. Beside NJ-MCL, the other four methods are very well known and therefore do not need any further introduction about them [28]. NJ-MCL method in brief; a new method has been developed which is a balance algorithm based on maximum likelihood (ML) and neighbor joining (NJ) algorithms. Algorithm of NJ-MCL method is based on the simultaneous estimation of all pairwise distances by maximizing a likelihood function and then NJ method is used to infer phylogeny [29]. The method of simultaneous estimation of pairwise distances called maximum composite likelihood (MCL) method. The NJ tree

[1] Accuracy: a phylogenetic method has high accuracy if it quickly converges on the true tree as more data are applied to the problem.

[2] Consistency: a phylogenetic method is consistent for an evolutionary model, if the method converges on the true tree as the data becomes infinite.

***Corresponding author:** Anup Som, Center of Bioinformatics, Institute of Interdisciplinary Studies, University of Allahabad, Allahabad-211002, India, Tel: +91 532 2460027; Fax: +91 532 2461009; E-mail: som.anup@gmail.com

Received April 11, 2013; **Accepted** July 27, 2013; **Published** July 29, 2013

Citation: Som A (2013) Genome-Scale Approach and the Performance of Phylogenetic Methods. J Phylogen Evolution Biol 1: 116. doi:10.4172/2329-9002.1000116

reconstruction using MCL method of distance estimation is referred as NJ-MCL method [30].

To search for MP trees, the Subtree-Pruning-Regrafting (SPR) search algorithm was used. In an extensive computer simulation, Takahashi and Nei [31] showed that the SPR search algorithm as efficient as the extensive search algorithms such as max-mini branch-and-bound search, min-mini heuristic search, and close-neighbor-interchange heuristic search. However, only the branch-and-bound search is guaranteed to find all the MP trees, but it takes prohibitive amount of time if the number of sequences is large (>15) [32].

Computer simulations to generate datasets

Figure 1 shows the model tree topology selected for computer simulations. This phylogeny is based on an independent analysis of 16,397 aligned nucleotide positions which included 19 nuclear and three mitochondrial genes (a total of 20 individual genes; three mt-genes making one alignment) for 42 placental mammals [33]. To conduct a biologically realistic simulation, gene-specific evolutionary parameters were estimated from each gene sequence of 42 mammals using the TN93 model [34] of sequence evolution, allowing for a gamma-distribution of rates. The TN93 model was selected the best-fit model of nucleotide substitution using the MODELTEST program [35,36]. The evolutionary parameters (i.e. branch length, transition/transversion ratio and gamma parameter) of each gene dataset were calculated using PhyML [37,38]. Gene sequences were simulated using model topology (Figure 1) along with gene-specific evolutionary parameters which are extracted from real data. For each set of gene-specific evolutionary parameters 100 replicate datasets were generated using the Dawg program [39] under the TN93 model of nucleotide substitution with a gamma distribution of rates.

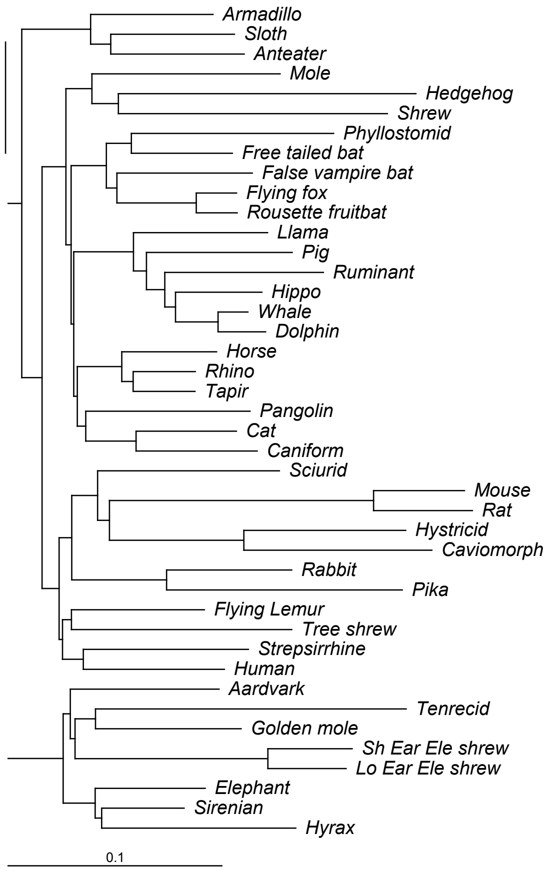

Figure 1: The model topology used in the computer simulations based on the 42-taxa tree from Springer et al. [33].

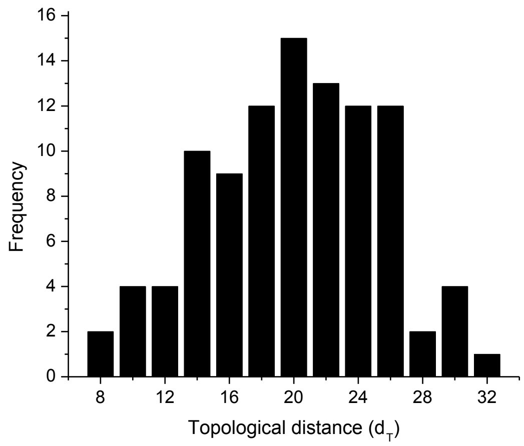

Figure 2: Distributions of the topological distances (d_T's) from the 100 replicates of EDG1 gene is plotted (978 bp). It shows that within replicates topological distances varied widely even when a unique set of evolutionary parameters are used in the computer simulation to generate replicates of EDG1 gene.

Phylogenetic analysis

Five methods of tree reconstruction were used in this study. Phylogenetic analyses were carried out using MEGA5 [30] for NJ-MCL, MP, and NJ methods, PhyML [38] for the ML method, and BAMBE [40] for the Bayesian method. The TN93 model of sequence evolution with gamma rate heterogeneity was used for Bayesian, ML and NJ methods. The MP trees were reconstructed using the (SPR) search algorithm [32]. The algorithm of NJ-MCL method was developed based on the TN93 model of sequence evolution. Therefore, for NJ-MCL method the TN93 model of sequence evolution (the default choice) with gamma rate heterogeneity was used. For a given simulated gene (whether containing an alignment of one simulated gene or the concatenation of multiple genes), Bayesian, ML, NJ-MCL, MP, and NJ trees were reconstructed and in each case the topological distances between the reconstructed trees and the model tree were estimated.

Accuracy of the inferred phylogeny

The accuracy of each method was calculated by the percentage of clades reconstructed correctly (P_C). This was obtained by $P_C = 100[1-d_T/(2m-6)]$, where d_T is the topological distance between the reconstructed and model trees and m is the number of sequences in the phylogeny [41,42]. All comparisons were made between the reconstructed trees and the model tree (Figure 1). For example, for a given simulated dataset, the Bayesian tree was reconstructed and then d_T was estimated between the reconstructed Bayesian tree and the model tree, and finally P_C was calculated from d_T value. A similar analysis is done for all five methods and also for each multigene dataset.

Construction of multigene datasets

For comparing the performance of the five different phylogenetic methods, 100 simulated datasets for each of 20 genes were generated. In construction of multigene datasets, one replicate should be selected out of the 100 replicates for each gene. To keep the replicate selection unbiased and realistic, the distribution pattern of d_T for 100 replicates of each simulated gene were plotted, and it was found that topological distances among the trees based on the replicates varied widely (Figure 2). Considering the nature of topological-distance distribution, the multigene datasets were reconstructed in three ways. These are (i) the best-replicate (BS) scenario where the simulated replicate for a given gene was selected that produced a phylogeny with the highest P_C (i.e., with the lowest topological difference when compared to the model tree); (ii) the worst-replicate scenario (WS) where the simulated replicate for a given gene was selected that produced a phylogeny with the lowest P_C; and (iii) the random-replicate (RS) scenario where all analyses were conducted using randomly chosen replicates to represent individual genes.

Progressive concatenation of genes

In this study relative performances of five different phylogenetic methods are investigated under the genome-scale approach. It is well known that adding a single gene to an existing set of genes improves the accuracy of phylogenetic reconstruction [9,43] and it is also obvious that more accurate method will converge first (i.e. will need a smaller number of genes to recover the true tree). The question is in which order the genes should be concatenated, because genes length vary widely and moreover they contain different levels of phylogenetic signal. Theoretically it is expected that concatenation of longer genes will converge first because longer genes get enough nucleotide substitutions and make it possible to infer them with greater accuracy. Moreover, an overall increase in sequence length would lead to reduce stochastic errors for evolutionary distances and other parameters in model based methods (i.e., Bayesian, ML, NJ-MCL, and NJ) [32]. Therefore, based on the variable length of genes along with different levels of phylogenetic signal three concatenation scenarios are considered: (i) longest to shortest (LS) genes concatenation where genes should be concatenated in the descending order of their length; (ii) shortest to longest (SL) genes concatenation where genes should be concatenated in the ascending order of their length; and (iii) random concatenation (RC) where genes have been selected randomly. The reasons for choosing these three concatenation scenarios are to examine the performance of the methods in a wide frame and examine which method takes a lower number of datasets to recover the model tree, to investigate whether concatenation of longest and smallest genes makes any effect on the methods performance, and to check if all three scenarios take almost same number of characters. Therefore, this study included all three concatenation scenarios (i.e., LS, SL and RC) under each of the three replicate selection scenarios (i.e., best-replicate, worst-replicate and random-replicate scenarios) [Supplementary data].

Results

Comparison of the efficiencies of phylogenetic methods

In this study genome-scale approach has been used in a biologically realistic computer simulation to examine the performance of the five major phylogenetic methods. Simulation results indicate that probabilistic methods (i.e., Bayesian and ML methods) are much more accurate than the NJ-MCL, MP and NJ methods and these results hold for each replicate scenario and for all three concatenation scenarios (Table 1). In general, the two probabilistic methods show more or less the same performance, but in a fine comparison ML is better than the Bayesian method (ML is more accurate and consistent than the Bayesian method). This analysis shows that the NJ-MCL, MP, and NJ methods are fast (i.e., computationally efficient), but their accuracy and consistency are very poor compared to the Bayesian and ML methods. Among NJ-MCL, MP and NJ methods, NJ-MCL is more accurate than MP and NJ with one exception; for the best replicate scenario, MP outperforms NJ-MCL and NJ methods (Figure 3a). In a further investigation, it was found that, for all three replication scenarios, Bayesian and ML methods take concatenation of few genes to recover the true tree (i.e., model tree) whereas other three methods take concatenation of considerably large number of genes (three to six times more number of genes were required by the NJ-MCL, MP and NJ methods) to recover the true tree. These results established the fact that the probabilistic methods (i.e., ML and Bayesian methods) are much more consistent and efficient than the NJ-MCL, MP and NJ methods. For example, under random-replicate scenario and for LS gene concatenation scenario Bayesian and ML methods take concatenation of 4 and 2 genes respectively to recover the true tree, whereas NJ-MCL, MP and NJ take concatenation of 7, 20, and 12 genes respectively to recover the true tree. These results are shown in Table 1.

More genes or more characters

In this study it was investigated how many characters are needed to recover the true tree and are they correlated with number of concatenated genes (i.e., more genes and also more characters or vice versa). Results in Table 1 show that, for best-replicate scenario, shortest to longest (SL) concatenation takes the less number of characters (with more genes) followed by LS and RC scenarios to recover the true tree, which is true for all five methods. For ML method under best-replicate scenario, it was found that LS concatenation takes 5812 characters (concatenation of 3 genes) whereas SL scenario takes only 1934 characters (concatenation of 6 genes) to recover the true tree. Similarly, for Bayesian method LS concatenation takes 5812 characters and SL concatenation takes 4042 (concatenation of 3 and 10 genes respectively). This result indicates that it is not always useful to consider longer genes; rather concatenation of smaller gene sequences, may be more number of genes with less number of characters, is more effective for reconstructing multigene phylogenies. Particularly, it should reduce the computational time. Moreover, this finding violates the theoretical expectation; concatenation of longer genes will converge first because longer genes get enough nucleotide substitutions and make it possible to infer them with greater accuracy. Although no one can guarantee

Method	Best replicate scenario						Random replicate scenario						Worst replicate scenario					
	LS		SL		RC		LS		SL		RC		LS		SL		RC	
	Gene	Char	Gene	Char	Gene	Char	Gene	Char	Gene	Char	Gene	Char	Gene	Char	Gene	Char	Gene	Char
Bayesian	3	5812	10	4042	9	8440	4	6988	14	7318	15	13301	5	8077	17	10585	12	11002
ML	3	5812	6	1934	8	7438	2	4564	14	7318	13	12649	2	4564	17	10585	12	11002
MCL	17	15548	17	10585	18	14945	7	10057	20	16397	18	14945	9	11665	19	13480	13	12649
MP	11	13032	16	9409	14	12986	20	16397	16	9409	20	16397	>20	--	>20	--	>20	--
NJ	17	15548	17	10585	19	16193	12	13593	20	16397	20	16397	7	10057	19	13480	19	16397

Table 1: The number of genes and characters required to recover the true tree for each concatenation scenario and for the Bayesian, ML, MP, NJ-MCL, and NJ methods. Columns "Gene" and "Char" stand for total number of genes and corresponding total number of characters used in a single concatenation to infer the true tree. LS, SL and RC concatenation scenarios represent longest to shortest, shortest to longest and random concatenation of genes respectively.

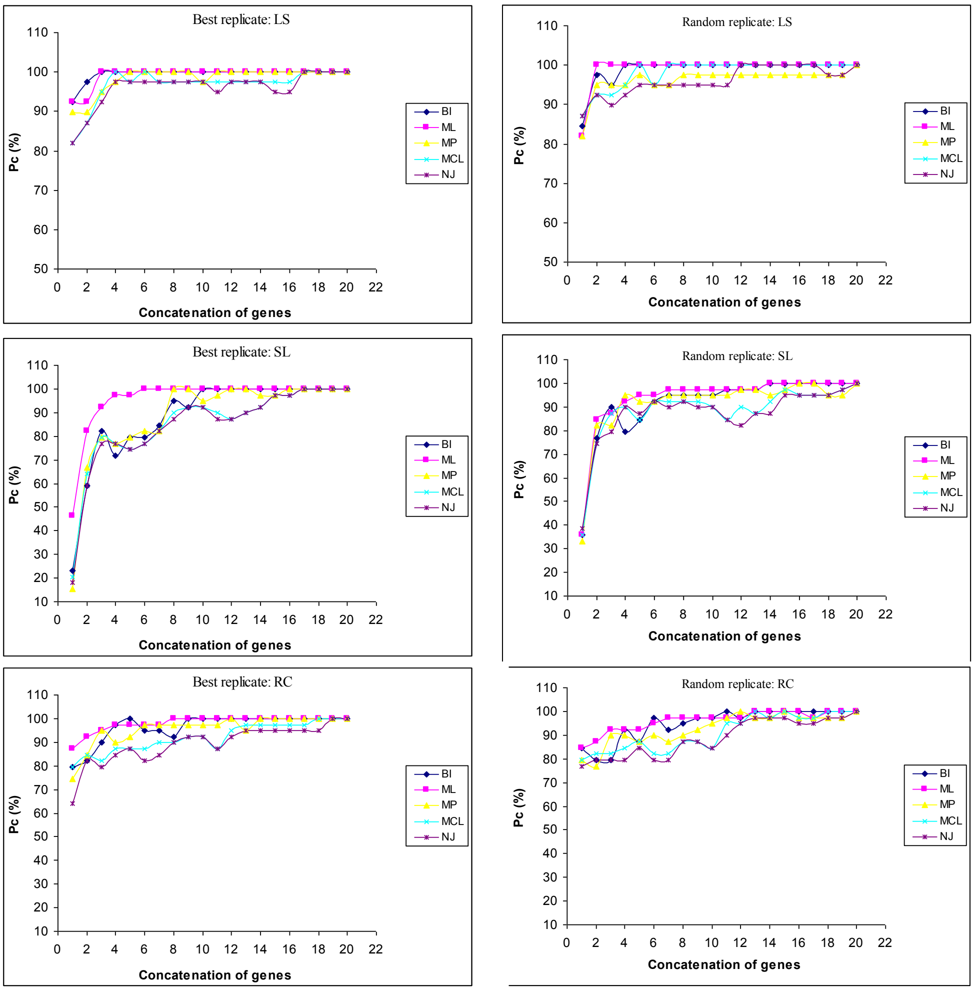

Figure 3: Percentage of branches inferred correctly (P_C) is plotted against concatenated genes for three concatenation approaches (i.e., longest to shortest (LS), shortest to longest (SL), and random concatenation (RC) approaches) and for (a) best-replicate and (b) random-replicate scenarios. Plots show that probabilistic methods (i.e., Bayesian & ML) converge (i.e., P_C=100%) much faster than the other methods (i.e., NJ-MCL, MP, & NJ). Furthermore in case of the ML method P_C value is increased (or remains unchanged) with the addition of a single gene to an existing concatenation, whereas for other methods, in several cases, addition of a gene produce more incorrect tree than that from initial dataset (see text).

that the concatenation of smaller genes will produce better phylogeny because the levels of phylogenetic signals present in the sequences are most important factors for reconstructing true evolutionary history of the species or genes. For other replicate scenarios (i.e., WS and RS) show SL case takes more gene and also more characters than LS concatenation. This contradiction is due to the quality of the gene replications.

Effect of addition of a single gene to an existing concatenation

It was also investigated how addition of a single gene to an existing concatenation (that generates a new concatenated dataset) improves the accuracy of the methods. Figure 3 shows the concatenated gene versus P_C plots for all three concatenation scenarios (i.e., for LS, SL and RC scenarios) and for best-replicate and random-replicate scenarios. In overall, progressive addition of genes improved the accuracy of the phylogenetic reconstruction for all five methods. However for NJ-MCL, MP, and NJ methods, in several cases (Figure 3), addition of a single gene to an existing concatenation decreases the P_C value obtained from initial dataset (i.e., addition of a gene produce more incorrect tree than that from initial dataset). This is due to different phylogenetic signal of the individual genes, either because of real differences in their evolutionary history, or because of different statistical biases, and NJ-MCL, MP, and NJ methods failed to accommodate such properties. In this situation concatenation may obscure the underlining species tree [44]. Interestingly, in spite of rigorous statistical properties Bayesian method also suffers from similar problem, but the performance is comparatively better than the NJ-MCL, MP, and NJ methods. On the other hand, in case of ML method addition of a single gene to an existing concatenation mostly improves the phylogenetic reconstruction or keeps its accuracy (P_C) as obtained from previous data (i. e., P_C value is increased with the addition of a gene or remain unchanged). This result states that the ML method is more consistent than all other methods.

How many datasets are needed to recover the true tree?

Another investigation was performed to find out how much datasets are needed to recover the true tree. The results showed that each different method takes different number of genes depending on their statistical power to resolve branches. Even for a particular method numbers of genes are varied among different replicate scenarios. Table 1 shows the results of such variations. For example, for the Bayesian

Phylogenetic method	Computer program	Time required
Bayesian	BAMBE	4 hr, 28 min, 48 sec
ML	PhyML	10 min, 40 sec
MCL	MEGA4	2 sec
MP	MEGA4	10 sec
NJ	MEGA4	1.5 sec

Table 2: Average run times for various methods. The computing times were measured on a PC Pentium IV 3.0 GHz (2 GB RAM) running with Windows XP. Datasets of 42 taxa with 16,397 bp were used to estimate the average computation time.

method under LS concatenation, the best, worst, and random replicates take three, four, and five genes respectively. These results imply that the quality of replication is a primary factor and in a simulation study it is possible to distinguish the best and worst replicates, but in reality it is not possible. This simulation experiments show the number of genes sufficient to recover the true tree ranged from a minimum of 4 to 20. This result completely agreed with Rokas et al. [43].

Discussion

In this article, relative performance of five major phylogenetic methods were evaluated under the genome-scale approach using biologically realistic simulated nucleotide data and simulation, and results show that the Bayesian and ML methods are much more accurate than the NJ-MCL, MP and NJ methods. These results agreed with other studies with an exception [20-25]. In Hall's study [22], Bayesian method is more accurate than the ML method. By contrast, this study shows ML method is slightly better than Bayesian method. This is apparently due to a difference of our simulations strategy and methodologies of the experiment. Beside comparison of the performance of methods this study has revealed several important aspects of genome-scale approach such as how concatenations of longest and smallest genes make effect on the methods performance, how much dataset are needed to recover the true tree, and whether more genes or more characters are important. These have been explained in the results section.

Concerning the accuracy of methods, the results showed that probabilistic methods (i.e., Bayesian and ML methods) are much more accurate than the NJ-MCL, MP and NJ methods. In overall, ML is more accurate, followed by the Bayesian, NJ-MCL, NJ, and MP methods. Furthermore, it has been shown that the ML method is much more consistent than other methods (even superior to the Bayesian method). An accurate algorithm may be useless if it is too slow. Therefore, for comparison proposes, the run time of each algorithm was measured which is shown in Table 2. Although NJ-MCL, MP, and NJ methods are very much computational efficient, but their accuracies and consistencies are very poor compared to the Bayesian and ML methods. On the other hand, Bayesian is very efficient, but less consistent and takes much longer execution time; whereas ML is very accurate, consistent, and computational efficient. Therefore, in conclusion, the continued preference of the ML method is recommended when genome-scale approach is used for phylogenetic reconstructions.

Acknowledgements

I acknowledge useful discussions with Dr. Dan Graur and Dr. Giddy Landan of University of Houston. I thank Dr. Georg Fuellen of University of Rostock for his helpful comments. I also thank Priyanka Sengupta for editorial support.

Supplementary Data

Supplementary data associated with this article can be found in the online version.

References

1. Erdos PL, Steel MA, Szekely LA, Warnow TJ (1999) A few logs suffice to build (almost) all trees (I). Random Struct. Algorithms 14: 153-184.
2. Bininda-Emonds OR, Brady SG, Kim J, Sanderson MJ (2001) Scaling of accuracy in extremely large phylogenetic trees. Pac Symp Biocomput 2001:547-58.
3. Qiu YL, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, et al. (1999) The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature 402: 404-407.
4. Soltis PS, Soltis DE, Chase MW (1999) Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402: 402-404.
5. Graham SW, Olmstead RG (2000) Utility of 17 chloroplast genes for inferring the phylogeny of the basal angiosperms. Am J Bot 87: 1712-1730.
6. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ (2001) Universal trees based on large combined protein sequence data sets. Nat Genet 28: 281-285.
7. Madsen O, Scally M, Douady CJ, Kao DJ, DeBry RW, et al. (2001) Parallel adaptive radiations in two major clades of placental mammals. Nature 409: 610-614.
8. Murphy WJ, Eizirik E, Johnson WE, Zhang YP, Ryder OA, et al. (2001) Molecular phylogenetics and the origins of placental mammals. Nature 409: 614-618.
9. Bapteste E, Brinkmann H, Lee JA, Moore DV, Sensen CW, et al. (2002) The analysis of 100 genes supports the grouping of three highly divergent amoebae: Dictyostelium, Entamoeba, and Mastigamoeba. Proc Natl Acad Sci U S A 99: 1414-1419.
10. Gontcharov AA, Marin B, Melkonian M (2004) Are combined analyses better than single gene phylogenies? A case study using SSU rDNA and rbcL sequence comparisons in the Zygnematophyceae (Streptophyta). Mol Biol Evol 21: 612-624.
11. Delsuc F, Brinkmann H, Philippe H (2005) Phylogenomics and the reconstruction of the tree of life. Nat Rev Genet 6: 361-375.
12. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311: 1283-1287.
13. Philippe H, Lopez P (2001) On the conservation of protein sequences in evolution. Trends Biochem Sci 26: 414-416.
14. Felsenstein J (1978) Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology 27: 401-410.
15. Kolaczkowski B, Thornton JW (2004) Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. Nature 431: 980-984.
16. Philippe H, Zhou Y, Brinkmann H, Rodrigue N, Delsuc F (2005) Heterotachy and long-branch attraction in phylogenetics. BMC Evol Biol 5: 50.
17. Jeffroy O, Brinkmann H, Delsuc F, Philippe H (2006) Phylogenomics: the beginning of incongruence? Trends Genet 22: 225-231.
18. Som A, Fuellen G (2009) The effect of heterotachy in multigene analysis using the neighbor joining method. Mol Phylogenet Evol 52: 846-851.
19. Queiroz AD (1993) For consensus (Sometimes). Syst Biol 42: 368-372.
20. Saitou N, Imanishi T (1989) Relative efficiencies of the Fitch-Margoliash, maximum parsimony, maximum-likelihood, minimum-evolution, and Neighbor-Joining methods of phylogenetic tree construction in obtaining the correct tree. Mol Biol Evol 6: 514-525.
21. Kuhner MK, Felsenstein J (1994) A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol Biol Evol 11: 459-468.
22. Hall BG (2005) Comparison of the accuracies of several phylogenetic methods using protein and DNA sequences. Mol Biol Evol 22: 792-802.
23. Huelsenbeck JP (1995) Performance of phylogenetic methods in simulation. Syst Biol 44: 17-48.
24. Som A (2009) ML or NJ-MCL? A comparison between two robust phylogenetic methods. Comput Biol Chem 33: 373-378.
25. Yang Z, Rannala B (2012) Molecular phylogenetics: principles and practice. Nat Rev Genet 13: 303-314.
26. Penny D, Hendy MD, Steel MA (1992) Progress with methods for constructing evolutionary trees. Trends Ecol Evol 7: 73-79.

27. Hillis DM (1995) Approaches for assessing phylogenetic accuracy. Syst Biol 44:3-16.

28. Felsenstein J (2004) Inferring phylogenies. Singular Associates, Sunderland, Massachusetts.

29. Tamura K, Nei M, Kumar S (2004) Prospects for inferring very large phylogenies by using the neighbor-joining method. Proc Natl Acad Sci U S A 101: 11030-11035.

30. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.

31. Takahashi K, Nei M (2000) Efficiencies of fast algorithms of phylogenetic inference under the criteria of maximum parsimony, minimum evolution, and maximum likelihood when a large number of sequences are used. Mol Biol Evol 17:1251-1258.

32. Nei M, Kumar S (2000) Molecular Evolution and Phylogenetics. Oxford University Press: New York.

33. Springer MS, Murphy WJ, Eizirik E, O'Brien SJ (2003) Placental mammal diversification and the Cretaceous-Tertiary boundary. Proc Natl Acad Sci U S A 100: 1056-1061.

34. Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512-526.

35. Posada D, Crandall KA (1998) MODELTEST: testing the model of DNA substitution. Bioinformatics 14: 817-818.

36. Posada D, Crandall KA (2001) Selecting the best-fit model of nucleotide substitution. Syst Biol 50: 580-601.

37. Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, et al. (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 59: 307-321.

38. Guindon S, Lethiec F, Duroux P, Gascuel O (2005) PHYML Online--a web server for fast maximum likelihood-based phylogenetic inference. Nucleic Acids Res 33: W557-559.

39. Cartwright RA (2005) DNA assembly with gaps (Dawg): simulating sequence evolution. Bioinformatics 21 Suppl 3: iii31-38.

40. Simon, D and B Larget (2000) Bayesian analysis in molecular biology and evolution (BAMBE), version 2.03 beta. Department of Mathematics and Computer Science: Duquesne University.

41. Robinson DF, Foulds LR (1981) Comparison of phylogenetic trees. Math Biosci 53: 131-147.

42. Penny D, and Hendy MD (1985) The use of tree comparison metrics. Syst Zool 34: 75-82.

43. Rokas A, Williams BL, King N, Carroll SB (2003) Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425: 798-804.

44. Phillips MJ, Delsuc F, Penny D (2004) Genome-scale phylogeny and the detection of systematic biases. Mol Biol Evol 21: 1455-1458.