Inference and Applications of Molecular Phylogenies: An Introductory Guide

Korbinian Strimmer and David L. Robertson

Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK.

December 5, 2000; revised July 17, 2001

Corresponding author:

Korbinian Strimmer

Voice: +44-1865-271272

Fax: +44-1865-271249

Email: korbinian.strimmer@zoo.ox.ac.uk

Number of figures: 1

Number of tables: 2

Number of formulas: 0

Number of boxes: 2

Number of pages (main text): 13

Number of references: 92

1 Introduction

Molecular phylogenies have a wide range of practical applications in the analysis of DNA sequences and are now an essential tool in areas ranging from population genetics to genomics to virology. Modern computers have fostered the development of sophisticated methodologies, and subsequently a large number of programs have become available. In this chapter we give an introductory overview of the most important methods of inferring phylogenetic trees from nucleotide or amino acid sequence data, emphasizing concept rather than mathematical detail (Box 1). We discuss simple guidelines for choosing a method appropriate for a data set. Advanced issues in molecular phylogenetics are briefly introduced, along with suggestions for further reading. The versatility of molecular phylogenetics is also highlighted and we present a range of practical problems where evolutionary trees have formed a key component of the analysis. Finally, we provide an overview of phylogenetic software and corresponding Internet resources (Box 2).

2 Inference of Phylogenetic Trees

A natural means to illustrate the evolutionary relationships among a sample of DNA sequences is in the form of a tree (Figure 1). The branching order of the tree (its topology) indicates how the sequences are related to each other, i.e. which sequences share a most recent common ancestor. If included, branch lengths of a tree represent genetic distance. Evolutionary trees can be depicted as *unrooted* (Figure 1a and c) or *rooted* (Figure 1b).

Whenever sequence data are analyzed it is important to account for the statistical dependencies between the sequences indicated by the phylogenetic tree. However, true evolutionary relationships among the sequences are rarely known. Instead, the

phylogenetic tree is inferred from the data. Unfortunately, reconstructing accurate evolutionary trees from genetic data is intrinsically difficult, due to three main factors:

First, to infer a tree from DNA sequences a reliable multiple alignment must be constructed. However, an alignment in itself infers a particular evolutionary history by attempting to determine which sites along the sequences are homologous. Moreover, if the alignment is unreliable the accuracy of any tree inferred from it will be compromised.

Second, the number of possible trees relating a set of DNA sequences increases very rapidly as the number of sequences increases. Table 1 show the number of possible rooted and unrooted tree topologies for 3 to 50 sequences. Even for moderate numbers of sequences the number of possible trees is extremely large, rendering it impossible to scrutinize all possible arrangements in reasonable computer time.

Third, evolutionary relationships can be very complex, and similarity between sequences is generally not a good indicator of phylogenetic relatedness. For example, in Figure 1a the shortest genetic distance is between sequences A and C, while A and B are apparently more distantly related. However, when the tree is rooted with a known outgroup (sequence D in Figure 1b) it becomes clear that sequences A and B (and not A and C) share a more recent common ancestor. The same can be observed in the unrooted tree in Figure 1c where again the genetic distance between A and C is smallest but A is nevertheless grouped with B. Therefore, except under special circumstances, simple cluster methods that rely purely on genetic similarity are not suitable for recovering evolutionary relationships.

For the above reasons, no general multi-purpose algorithm to infer evolutionary trees is known that is suitable for all kinds of data. Instead, a whole suite of complementary phylogenetic methods are commonly used, each with their particular strengths (and weaknesses). Generally the more rigorous a particular method the more computational resources it requires.

3

2.1 Important Methods

Exhaustive studies have been performed on simulated data to investigate the accuracy and other properties of tree reconstruction methods (1, 2, 3, 4). In the following we briefly introduce the three major classes of phylogenetic inference methods that are frequently used.

2.1.1 Maximum-Parsimony

Parsimony methods were among the first methods for reconstructing phylogenetic trees, and are still widely used. The philosophy behind this approach is "Occam's razor" applied to the distribution of substitutions along the branches of a tree. *Maximum-parsimony* aims at finding the tree that explains the observed pattern of nucleotides (or amino acids) with the smallest number of mutations possible (5). This is done by placing substitutions on possible tree topologies so as to minimize the total number of substitutions required to explain the nucleotide (or amino acid) at each branch tip. The most parsimonious reconstruction, i.e. the one requiring the fewest substitutions, is chosen for each site. The total number of evolutionary changes on a tree is the sum of the number of changes for each site, and the tree topology that has the fewest changes is chosen as the most parsimonious tree. The various types of parsimony methods (6) differ only in the particular details of how the above steps are pursued.

The maximum-parsimony approach generally works quite well. However, it also has a number of pitfalls that may at times render it impractical. First, the number of most-parsimonious trees can be quite large so that there can exist large numbers of equally valid solutions. Second, as an exhaustive tree search, i.e. the evaluation of all possible trees, is not feasible except for small data sets, generally no guarantee can be given that the optimal tree has been found. Third, the most-parsimonious tree is not

4

necessarily a good explanation for the data, particularly when sequences are short (7). Fourth, maximum-parsimony does not make efficient use of all the available data because constant and other non-parsimony-informative sites are ignored (6). Fifth, parsimony is particularly prone to the "long branches attract" problem, i.e. it tends to group long branches together even if this does not reflect the true evolutionary history (8).

2.1.2 Distance Methods

Distance-based methods work on pairwise genetic distances which have been computed from a sequence alignment, using a suitable *nucleotide or amino acid substitution model*, rather than analyzing the sites in the alignment directly.

Neighbor-joining (9, 10) is one of the most popular distance methods, primarily due to its speed at finding a single "best" tree with large data sets. This clustering approach starts from a star-like tree and resolves the tree by iteratively joining groups of sequences together, and in each step minimizing the total sum of branch lengths. Neighbor-joining has been shown to be efficient at recovering the correct tree topology (3, 4).

In contrast, *UPGMA* (11) simply clusters groups of sequences that exhibit the smallest distance between each other. As phylogenetic relationships are not necessarily linked to sequence similarity, this approach often fails to reconstruct the true tree except when the sequences have evolved in a clock-like manner, i.e. requires that the rate of evolution has been constant (3).

The *least-squares* method (12, 13) optimizes branch lengths on a tree by minimizing the error of the tree-induced pairwise distances when compared to the distances computed from the alignment. An optimal tree is then found by evaluating all possible (or all feasible) candidate trees. The least-squares approach has a solid

statistical foundation and so is often considered the best distance method (14).

Minimum-evolution (15), in a way similar to the least-squares method, also uses a tree search to find the best tree. The selection criterion in this approach is the total sum of branch lengths. Thus, it can be considered as somewhat similar to maximum-parsimony but applied to distance data. Note that a neighbor-joining tree is an approximation of the minimum-evolution tree.

2.1.3   Maximum-Likelihood

Maximum-likelihood (16) is the approach that is generally considered to make the most efficient use of the data and to provide the most accurate estimates of a phylogenetic tree. It is, however, also the most computationally demanding technique for reconstructing phylogenetic trees.

The basic idea of the likelihood approach is to compute the probability of the observed data assuming it has evolved under a particular evolutionary tree and a given probabilistic model of nucleotide (or amino acid) substitution. The maximum-likelihood (ML) tree is then the tree with the highest probability of explaining the data. Technically, the likelihood of a tree is computed using so-called directed graphical model for the sequence data (16, 17).

Inferring evolutionary trees using the maximum-likelihood principle is difficult for three main reasons. First, the likelihood itself is complicated to compute, essentially due to the problem of inferring the unknown ancestral character states in a tree. Second, even for a single tree topology, estimation of branch lengths is a very difficult optimization problem. Finally, computing likelihoods for all possible tree topologies is an impossible task even for medium-sized sets of sequences so that instead a heuristic tree search must be used. Thus it is no surprise that maximum-likelihood tree inference has only become popular since the widespread availability of faster

computers.

As a statistical approach the maximum-likelihood method enjoys a number of advantages, such as a generally lower variance than other estimators. Moreover, the likelihood framework provides a rigorous basis for statistical testing of competing hypothesis (18). In addition, it can be shown that maximum-likelihood encompasses the maximum-parsimony method, the latter being an approximation to maximum-likelihood (19). The maximum-likelihood approach can also be used to compute genetic distances between sequences, e.g. as prerequisite for a distance-based approach to tree reconstruction.

2.2   Common Features

To obtain a better understanding of the relationships between the discussed methods it is helpful to look at alternative classifications of phylogeny inference procedures. In Table 2 we present a number of criteria that highlight some shared features of the various methods.

An important characteristic is how the underlying data, i.e. the sequence alignment, is used. There are two main classes: methods based on character state data where sites in an alignment are individually analyzed, and methods based on distance data where a pairwise distance matrix relating all possible pairwise sequence comparisons to each other is used to summarize the information in the alignment. Methods that work directly on sequence data (e.g. maximum-likelihood) can be expected to be more accurate than those using distance data (e.g. least-squares). However, pairwise distances often provide a remarkably compact condensation of the major features of the data.

A second criterion which can be used to distinguish methods is the way the optimal tree is obtained. Some methods simply cluster sequences together according to a set

of given rules (e.g. neighbor-joining) whereas tree-search methods use specific optimality criteria to choose among all possible tree topologies. The latter methods that involve tree evaluation (specifically maximum-parsimony, maximum-likelihood, least-squares and minimum-evolution) can be implemented with a number of search strategies such as exhaustive search or branch-and-bound, which guarantee an optimal solution but also have potentially very large (!) run-times. Alternatively, heuristic searches based on rearrangement of subtrees (6) or reassembling of quartet trees (20, 21) can be employed but these do not guarantee the optimal tree will be found. Stochastic tree searches based, e.g., on genetic algorithms (22) or Markov chain Monte Carlo techniques (23) have also been implemented.

Another distinctive feature of phylogenetic tree reconstruction method is use of the model of substitution. The purpose of a substitution model is to estimate the actual amount of evolutionary change. This is problematic because there is not a simple relationship between the observed and actual changes, and the same site in a sequence might have undergone repeated substitutions while only the last substitution can be observed. Thus, any substitution model must "correct" for these unobserved changes. In nucleotide substitution models factors such as transitions being more frequent than transversions and unequal base composition are also incorporated into the model. For detailed discussion of the available models see ref. (6). Note that in maximum-parsimony the model of substitution is implicit in the method. This does not imply that maximum parsimony is model-free. Rather, methods like maximum-parsimony may often be inadequate for sequence data because of its simple implicit substitution model, whereas in maximum-likelihood and distance methods a complex, and hopefully realistic, model can be employed. Maximum-parsimony procedures have been implemented that attempt to give different weights to different types of substitutions, for example to transitions and transversions (6).

2.3 Selection of Suitable Method

When confronted with the large variety of methods available for analyzing a sequence data set it can often be difficult to choose a suitable method. However, the following simple guidelines may be useful:

1. If there are only a few sequences (< 10) in the data set then maximum-likelihood, using an exhaustive tree search, should be the method of choice. For a medium-sized data set (10-100 sequences) maximum-likelihood is still applicable, but heuristic tree searches have to be employed. Maximum-parsimony or least-squares are also good alternatives for this size of data set. For very large numbers of sequences (> 100) distance-based methods like neighbor-joining are the only computationally feasible approaches that can be used. It is important to keep in mind that the number of sites in the alignment will also impact the runtime of a program.

2. If at all possible more than one tree reconstruction method should be employed to check whether the same branching pattern is consistently recovered.

3. Note that if sequences are very similar or highly divergent the phylogenetic signal will be weak so it is generally unlikely that the corresponding tree can be fully resolved.

4. When selecting a model of substitution in maximum-likelihood, or with distance methods, it is a good idea to start with a simple substitution model, and to repeat the analysis with a more complex model and to carefully observe changes in tree topology and branch lengths.

2.4   Other Methods

The search for new phylogenetic methodologies and improved models of sequence evolution is a very active area of research and a wide range of phylogenetic methods have been proposed, many of which are variants of the methods mentioned before. For example, several modifications of neighbor-joining have been suggested (24, 25, 26). Research into methodologies has also focused on the more theoretical aspects of phylogenetic reconstruction, such as phylogenetic invariants (27, 28, 29, 30), phylogenetic spectra (31, 32), and phylogenetic geometry (33) and on the employment of advanced techniques like neural networks (34) or genetic algorithms (22). Much effort has also gone into developing sophisticated statistical models of substitution processes (35, 36, 37). Another research direction is the study of algorithms suitable for very large data sets (38, 39).

2.5   Phylogenetic Uncertainty

So far we have given a brief overview of how a phylogenetic tree can be estimated from molecular data. In the investigation of real data this step rarely is the end of the analysis, rather often it is only the beginning.

One of the most important questions is the problem of assessing the accuracy of the obtained tree. The most commonly used method is the bootstrap approach which has been developed to evaluate the stability of internal branches in trees (40, 41). Bootstrapping repeatedly samples sites with replacement from the original alignment until a new alignment of the same length as the original is obtained. The same phylogeny method used to construct the tree from the original alignment is then applied on the sampled alignment. This procedure is repeated a set number of times (commonly 1000) and if a particular grouping of sequences is found in 75% or more of replicates those clusters are considered well-supported (42). Other methods exist

10

that assess the phylogenetic signal of a data set prior to the inference of a tree (43).

A single "best" tree is only a point estimate of the true evolutionary relationships. Thus, the aim to better account for phylogenetic uncertainty has led to a general trend towards multiple-tree analysis. For example, likelihood ratio tests are frequently used to compare competing phylogenetic hypotheses in the form of alternative tree topologies (44). For many data sets it is, however, more suitable to introduce the notion of confidence sets of trees, where an ensemble of trees is preferred over a single tree. The likelihood method and the related Bayesian framework (23) are particularly amenable to this kind of analysis. This is an active area of research and various tests have been developed to determine confidence sets of trees (45, 46).

Closely related to this topic are evolutionary networks which are used to represent non-tree-like evolution, i.e. those relationships in a tree which cannot be represented by a tree. A large variety of different types of networks exists. As for evolutionary trees, there are parsimony (47, 48), distance-based (49) and likelihood methods (17, 50) for inferring evolutionary networks. Splitgraphs (51) merely represent incompatibilities in distance data while median networks (48) and ancestral recombination graphs (52, 53, 54) can be interpreted as collections of trees. A median network contains all the most-parsimonious trees for a data set, and an ancestral recombination graph with r recombination events combines $r + 1$ site-specific clock-like evolutionary histories.

2.6   Further Reading

General text-book introductions to molecular evolution and phylogenetics are provided by refs. (55, 56, 57). An introductory guide with detail on the use of some specific phylogeny software is given in ref. (58). Precise algorithmic details of tree reconstruction methods are reviewed in ref. (6) whereas ref. (59) gives a general

11

overview of probabilistic sequence analysis. Advanced theoretical topics in phylogeny are discussed in ref. (60).

3 Applications

Phylogenetic trees play an important role in a large variety of problems, from all fields of biology and related disciplines (61). The richness of applications illustrates the central position that molecular phylogeny has assumed in modern biology. In the following we list some interesting case studies. For further details we refer to the original papers.

3.1 Sequence Classification

Trees provide natural hierarchical classifications of the investigated sequences. Consequently, this has been one of the first applications of phylogenetic trees.

1. In taxonomy, species phylogenies are now commonly inferred using reconstructed gene trees. This is particularly interesting when there are no morphological data, or when molecular and morphological data contradict. For example, the relationship between the major taxonomic kingdoms has been established entirely from molecular sequences (62). In the inference of gene trees it is important to distinguish between the actual species tree and the embedded gene trees (63, 64).

2. In virology, phylogenetic trees have been used to define significant phylogenetic clusters of viruses. These can be used to test whether patients are linked epidemiologically and to study infected populations (65).

3. The prediction of gene function can be substantially refined by taking the phylogeny of the investigated gene into account (66). This approach is called phylogenomics (67).

3.2 Analysis of Population History

Phylogenetic trees also contain information about the demographic history of the population from which the sequences were sampled. This is relevant for problems in anthropology, epidemiology and virology.

1. Population size changes over time influence the shape and the branch lengths of trees. Coalescent theory (68, 69) provides a probabilistic model for this process and thus allows the extraction of information concerning population history from an inferred tree (70, 71).

2. Trees also allow the determination of the geographic origin of a population. For example, the common ancestor of human mitochondrial DNA (mitochondrial Eve) has been shown to be of African origin (72).

3. In addition, phylogenetic trees provide information about the time-frame of evolutionary events, e.g. about the age of the mitochondrial Eve (72) or the time of the human-ape split (73).

3.3 Processes in Molecular and Genome Evolution

Evolutionary processes in genes and genomes leave phylogenetic signals in the sequence data. Consequently, molecular phylogenies are important tools, for instance, for:

1. estimating model parameters and substitution rates (37),

2. detecting recombination in the evolutionary history of sequences (74, 47, 75),

3. inferring duplication events in genomes (76, 77),

4. inferring gene rearrangements in and between genomes (78),

5. detecting adaptive evolution, e.g. change of function and selection (79, 80, 81), and

6. comparing host and parasite phylogenies, i.e. for inferring co-phylogenies (55).

3.4   Comparative Studies

Comparative analysis aims at establishing correlations between traits across taxa, e.g. between brain and body size (82). It is important to discriminate dependencies between the investigated traits that are introduced merely by evolution from those representing true correlation. Consequently, phylogenetic comparative methods have been developed that explicitly take the underlying evolutionary tree into account (82, 83). Phylogenetic uncertainty can also be incorporated in such an analysis (84).

3.5   Other Bioinformatics Applications

There are many other applications of evolutionary trees in bioinformatics and sequences analysis. For example:

1. Phylogenetic trees are important for the reconstruction of ancestral sequences (5, 85, 86).

2. To compute sequence alignments some algorithms rely on the inference of evolutionary trees (59, 87).

3. Database searches for homologous sequence relationships can be made more efficient if the phylogeny of the sequences in the database is used as guide tree (88).

References

1. Huelsenbeck, J. P & Hillis, D. M. (1993) *Syst. Biol.* **42**, 247–264.

2. Hillis, D. M. (1995) *Syst. Biol.* **44**, 3–16.

3. Huelsenbeck, J. P. (1995) *Syst. Biol.* **44**, 17–48.

4. Strimmer, K & von Haeseler, A. (1996) *Syst. Biol.* **45**, 516–523.

5. Fitch, W. M. (1971) *Syst. Zool.* **20**, 406–416.

6. Swofford, D. L, Olsen, G. J, Wadell, P. J, & Hillis, D. M. (1996) in *Molecular Systematics*, eds. Hillis, D. M, Moritz, C, & Mable, B. K. (Sinauer Associates, Sunderland, Massachusetts), pp. 407–514.

7. Nei, M, Kumar, S, & Takahashi, K. (1998) *Proc. Natl. Acad. Sci. USA* **95**, 12390–12397.

8. Felsenstein, J. (1978) *Syst. Zool.* **27**, 401–410.

9. Saitou, N & Nei, M. (1987) *Mol. Biol. Evol.* **4**, 406–425.

10. Studier, J. A & Keppler, K. J. (1988) *Mol. Biol. Evol.* **5**, 729–731.

11. Sokal, R. R & Michener, C. D. (1958) *U. Kansas Sci. B.* **38**, 1409–1437.

12. Fitch, W. M & Margoliash, E. (1967) *Science* **155**, 279–284.

13. Cavalli-Sforza, L & Edwards, A. W. F. (1967) *Evolution* **21**, 550–570.

14. Felsenstein, J. (1997) *Syst. Biol.* **46**, 101–111.

15. Edwards, A. W. F. (1996) *Syst. Biol.* **45**, 79–91.

16. Felsenstein, J. (1981) *J. Mol. Evol.* **17**, 368–76.

17. Strimmer, K & Moulton, V. (2000) *Mol. Biol. Evol.* **17**, 875–881.

18. Goldman, N. (1993) *J. Mol. Evol.* **36**, 182–198.

19. Steel, M & Penny, D. (2000) *Mol. Biol. Evol.* **17**, 839–850.

20. Strimmer, K & von Haeseler, A. (1996) *Mol. Biol. Evol.* **13**, 964–969.

21. Strimmer, K, Goldman, N, & von Haeseler, A. (1997) *Mol. Biol. Evol.* **14**, 210–211.

22. Lewis, P. O. (1998) *Mol. Biol. Evol.* **15**, 277–283.

23. Larget, B & Simon, D. L. (1999) *Mol. Biol. Evol.* **16**, 750–759.

24. Gascuel, O. (1997) *Mol. Biol. Evol.* **14**, 685–695.

25. Bruno, W. J, Socci, N. D, & Halpern, A. L. (2000) *Mol. Biol. Evol.* **17**, 189–197.

26. Ota, S & Li, W.-H. (2000) *Mol. Biol. Evol.* **17**, 1401–1409.

27. Cavender, J. A. (1978) *Math. Biosci.* **40**, 271–280.

28. Cavender, J. A & Felsenstein, J. (1987) *J. Classif.* **4**, 57–71.

29. Felsenstein, J. (1991) *J. Theor. Biol.* **152**, 357–376.

30. Steel, M. (1994) *Appl. Math. Letters* **7**, 19–23.

31. Hendy, M. D & Penny, D. (1993) *J. Classif.* **10**, 5–24.

32. Hendy, M. D, Penny, D, & Steel, M. A. (1994) *Proc. Natl. Acad. Sci. USA* **91**, 3339–3343.

33. Kim, J. (2000) *Mol. Phylogenet. Evol.* **17**, 58–75.

34. Dopazo, J & Carazo, J. M. (1997) *J. Mol. Evol.* **44**, 226–233.

35. Goldman, N & Yang, Z. (1994) *Mol. Biol. Evol.* **11**, 725–736.

36. Yang, Z. (1996) *Trends Ecol. Evol.* **11**, 367–371.

37. Liò, P & Goldman, N. (1998) *Genome Res.* **8**, 1233–1244.

38. Csurös, M. (2001) *Proceedings of the 5th annual international conference on computational molecular biology (RECOMB)*, 104–113.

39. Kim, J. (1998) *Syst. Biol.* **47**, 43–60.

40. Felsenstein, J. (1985) *Evolution* **39**, 783–791.

41. Efron, B, Halloran, E, & Holmes, S. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 13429–13434.

42. Hillis, D. M & Bull, J. J. (1993) *Syst. Biol.* **42**, 182–192.

43. Strimmer, K & von Haeseler, A. (1997) *Proc. Natl. Acad. Sci. USA* **94**, 6815–6819.

44. Huelsenbeck, J. P & Rannala, B. (1997) *Science* **276**, 227–232.

45. Goldman, N, Anderson, J. P, & Rodrigo, A. G. (2000) *Syst. Biol.* **49**, 652–670.

46. Strimmer, K. (2001) *Proc. R. Soc. Lond. B submitted*.

47. Hein, J. (1993) *J. Mol. Evol.* **36**, 396–406.

48. Bandelt, H. J, Forster, P, Sykes, B. C, & Richards, M. B. (1995) *Genetics* **141**, 743–753.

49. Bandelt, H.-J & Dress, A. W. M. (1992) *Adv. Math.* **92**, 47–105.

50. Strimmer, K, Wiuf, C, & Moulton, V. (2001) *Mol. Biol. Evol.* **18**, 97–97.

51. Huson, D. H. (1998) *Bioinformatics* **14**, 68–73.

52. Hudson, R. R. (1983) *Theor. Popul. Biol.* **23**, 183–201.

53. Griffiths, R. C & Marjoram, P. (1996) *J. Comput. Biol.* **3**, 479–502.

54. Griffiths, R. C & Marjoram, P. (1997) in *Progress in Population Genetics and Human Evolution*, IMA Volumes in Mathematics and its Applications, eds. Donelly, P & Tavaré, S. (Springer Verlag, Berlin) Vol. 87, pp. 257–270.

55. Page, R. D. M & Holmes, E. C. (1998) *Molecular Evolution: A Phylogenetic Approach*. (Blackwell Science, Oxford).

56. Li, W.-H & Graur, D. (1999) *Fundamentals of Molecular Evolution*. (Sinauer Associates, Sunderland MA), 2nd edition.

57. Nei, M & Kumar, S. (2000) *Molecular Evolution and Phylogenetics*. (Oxford University Press, Oxford).

58. Hall, B. G. (2001) *Phylogenetic Trees Made Easy: A How-To Manual for Molecular Biologists*. (Sinauer Associates, Sunderland, Massachusetts).

59. Durbin, R, Eddy, S, Krogh, A, & Mitchison, G. (1998) *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. (Cambridge University Press, Cambridge).

60. Mirkin, B, McMorris, F. R, Roberts, F. S, & Rzhetsky, A, eds. (1997) *Mathematical Hierarchies and Biology: DIMACS Workshop November 13–15, 1996*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science. (American Mathematical Society, Providence, Rhode Island) Vol. 37.

61. Harvey, P. H, Leigh Brown, A. J, Maynard Smith, J, & Nee, S, eds. (1996) *New Uses for New Phylogenies*. (Oxford University Press, Oxford).

62. Doolittle, W. F. (1999) *Science* **284**, 2124–2128.

63. Maddison, W. P. (1996) *Syst. Biol.* **46**, 523–536.

64. Slowinski, J. B & Page, R. D. M. (1999) *Syst. Biol.* **48**, 814–825.

65. Crandall, K. A, ed. (1999) *The Evolution of HIV*. (John Hopkins University Press, Baltimore).

66. Pellegrini, M, Marcotte, E. M, Thompson, M. J, Eisenberg, D, & Yeates, T. (1999) *Proc. Natl. Acad. Sci. USA* **96**, 4285–4288.

67. Eisen, J. A. (1998) *Genome Research* **8**, 163–167.

68. Donnelly, P & Tavaré, S. (1995) *Annu. Rev. Genet.* **29**, 401–421.

69. Nordborg, M. (2001) in *Handbook of Statistical Genetics*, eds. Balding, D, Bishop, M, & Cannings, C. (Wiley, Chichester), pp. 179–212.

70. Pybus, O. G, Rambaut, A, & Harvey, P. H. (2000) *Genetics* **155**, 1429–1437.

71. Strimmer, K & Pybus, O. G. (2001) *Mol. Biol. Evol.* **?**, ?–?

72. Vigilant, L, Stoneking, M, Harpending, H, Hawkes, K, & Wilson, A. C. (1991) *Science* **253**, 1503–1507.

73. Hasegawa, M, Kishino, H, & Yano, K. (1985) *J. Mol. Evol.* **22**, 160–174.

74. Robertson, D. L, Sharp, P. M, McCutchan, F. E, & Hahn, B. H. (1995) *Nature* **374**, 124–126.

75. McGuire, G & Wright, F. (2000) *Bioinformatics* **16**, 130–134.

76. Eulenstein, O, Mirkin, B, & Vingron, M. (1997) in *Mathematical Hierarchies and Biology*, DIMACS Series in Discrete Mathematics and Theoretical Computer Science, in Mirkin et al. (60) pp. 71–93, pp. 71–93.

77. El-Mabrouk, N, Bryant, D, & Sankoff, D. (1999) in *RECOMB'99 Lyon-France*. (ACM), pp. 154–163.

78. Blanchette, M, Kunisawa, T, & Sankoff, D. (1999) *J. Mol. Evol.* **49**, 191–203.

79. Yang, Z. (1998) *Mol. Biol. Evol.* **15**, 568–573.

80. Suzuki, Y & Gojobori, T. (1999) *Mol. Biol. Evol.* **16**, 1315–1328.

81. Yang, Z & Bielawski, J. P. (2000) *Trends Ecol. Evol.* **15**, 496–503.

82. Harvey, P. H & Pagel, M. D. (1991) *The Comparative Method in Evolutionary Biology.* (Oxford University Press, Oxford).

83. Felsenstein, J. (1985) *Am. Naturalist* **125**, 1–12.

84. Huelsenbeck, J. P, Rannala, B, & Masly, J. P. (2000) *Science* **288**, 2349–2350.

85. Yang, Z, Kumar, S, & Nei, M. (1995) *Genetics* **141**. 1641–1650.

86. Pupko, T, Pe'er, I, Shamir, R, & Graur, D. (2000) *Mol. Biol. Evol.* **17**, 890–896.

87. Thompson, J. D, Higgins, D. G, & Gibson, T. J. (1994) *Nucleic Acids Res.* **22**, 4673–4680.

88. Rehmsmeier, M & Vingron, M. (1999) in *Proceedings of the 14th German Conference on Bioinformatics*. (Hannover), pp. 66–72.

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

90. Felsenstein, J. (1993) *PHYLIP: Phylogenetic Inference Package, version 3.5c.* (Department of Genetics, University of Washington, Seattle).

91. Adachi, J & Hasegawa, M. (1996) *MOLPHY: Programs for Molecular Phylogenetics, version 2.3.* (Institute of Statistical Mathematics, Tokyo).

92. Yang, Z. (2000) *Phylogenetic analysis by maximum-likelihood (PAML), version 3.0.* (University College, London).

Box 1: Steps for Inferring a Phylogenetic Tree

This chapter deals with methods (main text) and programs (Box 2) to infer phylogenetic trees from DNA sequences. The following are the basic steps for preparing the data for phylogenetic analysis and for obtaining a publication-ready output:

1. Retrieve homologous DNA or protein sequences, e.g. from a public database like NCBI's GenBank (`http://www.ncbi.nlm.nih.gov/`) or the EMBL Nucleotide Sequence Database (`http://www.ebi.ac.uk/embl/`).

2. Align sequences, either by hand using a sequence editor like SEAL (`http://evolve.zoo.ox.ac.uk`) or using a sequence alignment program, e.g. CLUSTAL X (`http://www-igbmc.u-strasbg.fr/BioInfo/ClustalX/`).

3. Reconstruct the tree and assess its accuracy (this chapter)

4. Make an illustration of the tree using, e.g., TreeEdit (`http://evolve.zoo.ox.ac.uk`). Other tree drawing programs are listed on the web page `http://evolution.genetics.washington.edu/phylip/software.html`.

Box 2: Internet Resources

A large number of software packages are available to infer evolutionary trees from sequence data. The variety of this software derives directly from the richness of the available methods, with many researchers providing implementations of their own and other people's methods. An exhaustive list of phylogenetic computer programs is maintained by Joseph Felsenstein at his web page `http://evolution.genetics.washington.edu/phylip/software.html`. Programs dealing with recombination in the ancestry of sequences are listed at D.L.R.'s web page `http://grinch.zoo.ox.ac.uk/RAP_links.html`.

Currently the most widely used program for inferring phylogenetic trees from nucleotide data (using parsimony, likelihood and distance methods) is PAUP* by David Swofford (89) (`http://www.lms.si.edu/PAUP`). PHYLIP (90) is an alternative collection of programs for the analysis of nucleotide and amino acid data created by Joseph Felsenstein (`http://evolution.genetics.washington.edu/phylip`). Phylo_win by Nicolas Galtier and Manolo Gouy again implements a variety of different methods (`http://pbil.univ-lyon1.fr`). Maximum-likelihood analysis for nucleotide and amino acid data is provided by MOLPHY (91) by Jun Adachi and Masami Hasegawa (`ftp://sunmh.ism.ac.jp/pub/molphy`) and also by TREE-PUZZLE (20) by Heiko A. Schmidt, K.S., Martin Vingron and Arndt von Haeseler at (`http://www.tree-puzzle.de`). Parsimony and minimum-evolution methods are implemented in MEGA (57) by Sudhir Kumar and Masatoshi Nei (`http://www.megasoftware.net`).

A large selection of evolutionary models including codon-based models is implemented in PAML (92) by Ziheng Yang (`http://abacus.gene.ucl.ac.uk/software/paml.html`). A variety of useful

phylogenetic specialist tools created by Andrew Rambaut are available from the web page `http://evolve.zoo.ox.ac.uk`. For Bayesian tree inference there are two packages, BAMBE (23) by Bret Larget (`http://www.mathcs.duq.edu/larget/bambe.html`) and MrBayes by John Huelsenbeck (`http://brahms.biology.rochester.edu/software.html`). An object-oriented Java library for molecular evolution and phylogenetics is maintained by Alexei Drummond and K.S. (`http://www.pal-project.org`).

Web-based servers for phylogenetic inferences are also available. A large selection of programs is offered, e.g., by the server of the Institut Pasteur, Paris (`http://bioweb.pasteur.fr/seqanal/phylogeny/intro-uk.html`).

Table 1. Number of possible tree topologies.

sequences	unrooted trees	rooted trees
3	1	15
4	15	105
5	105	945
6	945	10395
7	10395	135135
8	135135	2027025
9	2027025	34359425
10	34359425	654729075
15	2.13458×10^{14}	6.19028×10^{15}
20	8.20079×10^{21}	3.19831×10^{23}
50	2.75292×10^{76}	2.72539×10^{78}

Table 2. Features of tree reconstruction methods.

Method	Data		Tree Search		Substitution Model	
	Character	Distance	Exhaustive[a]	Clustering	Explicit	Implicit
Maximum-parsimony	√		√			√
Neigbor-joining		√		√	√[b]	
UPGMA		√		√	√[b]	
Least-squares		√	√		√[b]	
Minimum-evolution		√	√		√[b]	
Maximum-likelihood	√		√		√	

[a] in practice a suitable heuristic shortcut will often have to be used.

[b] if model is used to compute pairwise distances.

Figure 1: Unrooted (a, c) and rooted (b) phylogenetic trees for sequences A, B, C, and D. Branch lengths represent genetic distance (e.g. nucleotide substitutions).
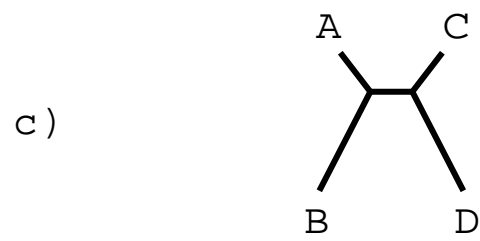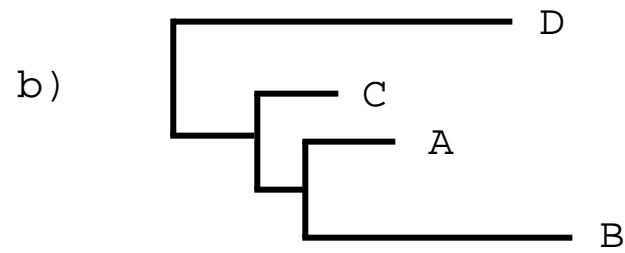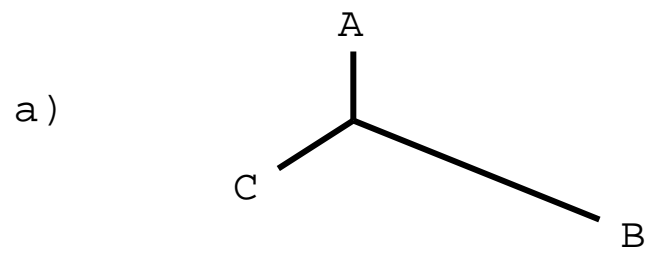
a)

A

C

B

b)

D

C

A

B

c)

A   C

B   D

Figure 1: