New Zealand Journal of Botany, 1993, Vol. 31: 307–316
0028–825X/93/3103–0307 $2.50/0

Inferring plant evolutionary history from molecular data

MICHAEL T. CLEGG

Department of Botany and Plant Sciences
University of California
Riverside, CA 92521–0124, U.S.A.

BRANDON S. GAUT

Department of Statistics
North Carolina State University
Raleigh, NC 27695, U.S.A.

MELVIN R. DUVALL
JOEL DAVIS

Department of Botany and Plant Sciences
University of California
Riverside, CA 92521–0124, U.S.A.

Abstract The rapid development of molecular methods during the 1980s has had a profound effect on the study of plant evolution. Molecular data have accumulated very rapidly, and this abundance of data poses new problems for data analysis. We illustrate three problem areas that arise in plant evolutionary inference. The first problem concerns the use of molecular data to analyse closely related plant species linked through reticulate evolution. The second problem arises from our ability to sample DNA sequences from alleles of a single genetic locus within plant species. We show how cumulative information on selection and random genetic drift can be extracted from such data. The third problem area concerns the limitations of current algorithms for phylogenetic inference when confronted with large sets of DNA sequence data. Based on a consideration of these problem areas, we conclude that: (1) asymmetric transmission of cpDNA markers is useful in resolving the parentage of hybrid plant taxa; (2) simple clustering algorithms can provide useful information on cultivar or variety relationships, despite intervarietal hybridisation, if genetic similarities are averaged over sufficient loci; (3) samples of complete DNA sequence data from plant nuclear genes can provide a new dimension of information on historical effective population sizes and on the mechanisms that generate allelic diversity; (4) analyses of the chloroplast gene *rbc*L, sampled from across the monocotyledon class, reveal large variation in relative rates of nucleotide substitution—these variations, in turn, have important consequences for phylogeny estimation algorithms; and (5) the combined use of algorithms like parsimony and maximum likelihood may represent a more efficient approach to the phylogenetic analysis of very large (>100) sets of DNA sequences.

Keywords monocot phylogeny; gene trees; reticulate evolution; coalescence theory

B93022
Received 18 February 1993; accepted 26 July 1993

INTRODUCTION

A principle objective of evolutionary biology is to reconstruct the history of organic life on earth. There are many sources of evidence about biological history including paleontological materials, comparative studies of morphology and development, and the pattern of sequence change in macromolecules. Major advances in the acquisition of DNA sequence data during the past decade have led to an explosion of work on molecular evolution. Studies of plant systematics and plant evolution have both been strongly affected by these developments. New problems of data analysis and inference have arisen as the molecular database has expanded. The purpose of this article is to explore several of these problem areas based on our experiences in the analysis of molecular data from plant species. The first issue that we will discuss concerns the problem of reticulate evolution in plant species. The second issue concerns the role of intragenic recombination in gene evolution and the use of coalescence models

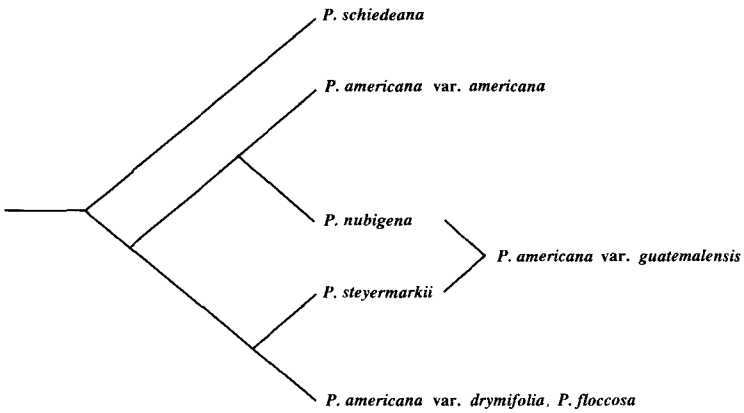

Fig. 1 A cladogram depicting the relationships among the varieties of cultivated avocado and closely related species of *Persea*. The variety *guatemalensis* is a hybrid with parents *P. steyermarkii* and *P. nubigena* (after Furnier et al. 1990).

to recover estimates of important population parameters. Our third topic concerns the computational problems associated with the analysis of very large sets of gene sequence data.

RETICULATE EVOLUTION IN AVOCADO

Owing to the importance of hybridisation among plant species, genetic relationships are sometimes represented by a complex network or reticulate pattern. The genetic history of cultivated avocado illustrates the complex patterns of reticulate evolution common to many long-lived woody plants. Avocado belongs to the genus *Persea*, itself an element of the Magnoliidae, which is among the oldest of dicot lineages. The cultivated avocado is assigned to the species *P. americana*, a member of the subgenus *Persea*. There are three botanical varieties of cultivated avocado: *P. americana* var. *americana* (West Indian variety denoted WI), var. *guatemalensis* (Guatemalan variety denoted G), and var. *drymifolia* (Mexican variety denoted M). The origins of domestication and the genealogical relationships among these varieties are obscure, although anthropological evidence suggests that avocado was under cultivation several thousand years ago (Smith 1966, 1969). There is also some debate about the taxonomic status of the varieties of cultivated avocado and of other closely related species of *Persea*. For instance, some workers have suggested that the species *P. nubigena* is actually a fourth variety of *P. americana* (Kopp 1966), while others have suggested that var. *guatemalensis* is a subspecies of *P. nubigena* (Bergh et al. 1973).

Furnier et al. (1990) employed a series of molecular markers to better characterise genetic relationships among the varieties of cultivated avocado and closely related species. Five species of *Persea* and 19 cultivars including all three varieties were studied using chloroplast DNA (cpDNA) and nuclear DNA (nDNA) probes. The probes included cDNA clones of the nuclear-encoded cellulase gene family, clones for the nuclear ribosomal RNA genes (rDNA), and a set of clones that spanned the entire cpDNA genome. These cloned probes were hybridised to Southern transfers of restriction-digested total avocado DNA. Eight restriction enzymes were used in the survey, and the basic data were numbers of mutational differences among pairs of entries as judged from changes in fragment mobility or RFLP patterns.

A fundamental consideration in evaluating the resulting data is the asymmetry of genetic transmission of cpDNA markers. In flowering plants, the predominant mode of cpDNA transmission is maternal, so that cpDNA-based phylogenies represent a network of maternal relationships. The nuclear markers are transmitted as classical Mendelian genes. The asymmetric transmission of the cpDNA markers revealed that G probably received its maternal genome from *P. steyermarkii* because two unique cpDNA markers linked G with *P. steyermarkii*. In addition, G was found to share a unique rDNA mutation with *P. nubigena*. The likely interpretation of these data is that G is an interspecific hybrid between *P. steyermarkii* and *P. nubigena*. Variety M was found to be most closely related to another species, *P. floccosa*. A cladogram depicting these relationships is shown in Fig. 1.

The study of Furnier et al. (1990) established several facts that were not obvious from the study of morphology and reproductive relationships. (1) The genetic data showed that the gene pool of

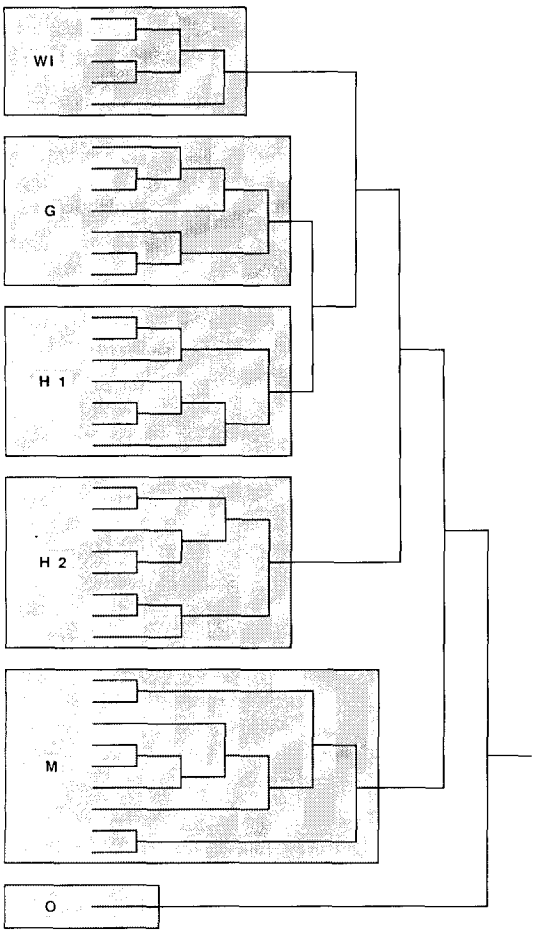

Fig. 2 A dendogram showing the results of a cluster analysis performed on a matrix of pairwise similarities (defined as the average fraction of genes shared in common between pairs of entries) based on 15 anonymous RFLP loci. The cluster WI is comprised of pure West Indian cultivars (var. *americana*), the cluster G is comprised of pure Guatemalan cultivars (var. *guatemalensis*), the cluster M is comprised of pure Mexican varieties (var. *drymifolia*), the cluster labelled O is an outgroup represented by *P. schiedeana*, the clusters labelled H1 and H2 include cultivars of known hybrid parentage and some cultivars of unknown parentage (but suspected to be hybrid).

cultivated avocado was quite broad and incorporated populations that had been assigned to separate species. (2) The data strongly suggested that the three varieties were domesticated separately. (3) The asymmetric transmission of the cpDNA markers established the hybrid origin of the G variety.

Avocado breeding has resulted in the naming of a number of cultivars, and the breeding histories of some of these cultivars are known. In others, the origin of the cultivars is obscure, owing to the common practice of selecting new cultivars from open-pollinated progenies where the paternal contribution is unknown. The cultivars are propagated asexually by bud grafting and, as a consequence, each named cultivar represents a single genotype. To further explore the use of genetic markers in establishing the breeding history of these materials, 15 anonymous clones that corresponded to single-copy DNA markers were hybridised to Southern transfers of restriction-digested DNA from 38 cultivars and to an outgroup species (*P. shiedeana*). A pairwise similarity measure was calculated as the fraction of genes shared in common at a locus, and the arithmetic average, taken over loci, of this measure was used to cluster cultivars according to an unweighted pair-group method algorithm (UPGMA) (Sneath & Sokal 1973). The resulting cluster analysis is shown in Fig. 2.

Cultivars representing each of the three varieties form distinct clusters. There are also two intermediate clusters that include cultivars that are derived from intervarietal hybrids. The breeding history of some of the hybrid cultivars is known; for example, the cultivar Hass (the predominant cultivar in commercial production in California) is a Gx(GxM) backcross. All of the known intervarietal hybrids, and their backcross derivatives, fall in the two intermediate clusters (denoted H1 and H2) and, in some cases, the particular cluster reflects the asymmetry of genetic contribution (e.g., Hass falls in the H1 cluster that is sister to G) consistent with the backcross origin of the Hass cultivar.

Our practical experience with avocado suggests that simple clustering algorithms give a useful picture of genetic relationships, even when the materials have a history of reticulate evolution, provided that averaging is done over a large number of genetic loci. This is not surprising because the average number of genes shared in common, when taken over a sufficient number of independent loci, should provide a good estimate of the proportionate contributions of the different hybridising lineages.

RECOMBINATION AND THE EVOLUTION OF THE *Adh1* LOCUS IN GRASS SPECIES

Sexually reproducing species are typically composed of a large number of lineages with complex historical interconnections, owing to the exchange of DNA

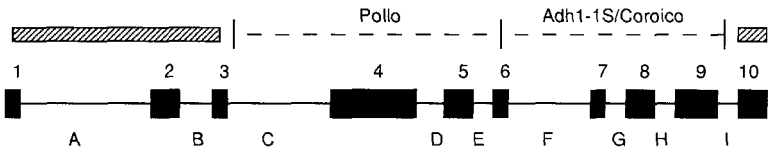

Fig. 3 Schematic diagram of the maize *Adh1*-C^m allele. Dark blocks (labelled 1–10) represent exons of the *Adh1*-C^m allele; lines connecting boxes (labelled A–I) represent intervening sequences. Parental alleles are represented by dashed lines. Hatching represents regions of uncertain origin in the *Adh1*-C^m allele.

segments through recombination among lineages. If attention is restricted to the DNA that encodes a particular gene, at some point in the past (T_0), all of the mutational forms of the gene will trace back to a single DNA copy (the coalescent). According to coalescence theory (Hudson 1991), the exact value of T_0 will be determined by the rate of mutation μ (per basepair, bp), the effective population size (N_e), and the pattern and strength of selection affecting the locus. In the absence of recombination, it is possible to estimate a gene phylogeny from samples of DNA sequences that reflect the history of mutational origin of the various alleles in the sample. The appeal of such data is that it provides a basis for inferring the cumulative importance of various populational processes, taken over a long period of time. This integration over time may provide a sensitive test for the relative importance of various patterns of selection versus genetic drift. To illustrate the kinds of inferences that can be drawn from samples of DNA sequences, we describe our studies of the alcohol dehydrogenase 1 locus (*Adh1*) in several grass species.

The enzyme alcohol dehydrogenase 1 (alcohol:NAD + oxidoreductase, EC 1.1.1.1) is important in the metabolic response to anaerobic stress in plants. The gene is expressed at a high level following flooding, and, as a consequence, considerable effort has been devoted to the molecular characterisation of the *Adh1* gene and its associated regulatory elements. The DNA sequence of *Adh1* has been reported from four grass species, and rates of molecular evolution have been estimated from these data (Gaut & Clegg 1991). The gene is approximately 3450 bp in length (from initiation of transcription to termination signals), and it is interrupted by nine introns of various lengths (see Fig. 3 for a diagram of the exon/intron structure).

Three genomic clones of *Adh1* representing different electromorphs have been sequenced from *Zea mays* subsp. *mays* (cultivated maize) (Dennis et al. 1984; Sachs et al. 1986; Osterman & Dennis 1989). To provide a larger sample of *Adh1* sequences, Gaut & Clegg (1993) determined the DNA sequence of 2098 bp of the gene from three maize landrace accessions and from two closely related species of perennial teosintes (*Z. diploperennis* and *Z. luxurians*). The additional sequence data begin in the third exon and continue to the tenth exon where satisfactory sites for PCR amplification were identified (Fig. 3). The total dataset permits the analysis of eight alleles over a 2098 bp region.

There are 88 polymorphic nucleotide sites among the eight maize alleles (approximately 4% of sites are polymorphic), of which 59 map into introns (59/1219) and 29 map into exons (29/879). Six of the exon sites represent amino acid polymorphisms, while the remainder are silent polymorphisms. Of the amino acid polymorphisms, one causes a charge change and has been identified as the site responsible for the fast/slow electromorph. Thus, maize *Adh1* is highly polymorphic at the DNA sequence level. Two of the eight alleles appear to be the products of intracistronic recombination (recombination between different *Adh1* alleles). One allele (*Adh1*-C^m) may represent three separate recombination events. Two of the parental alleles of *Adh1*-C^m are also present in the sample of just eight alleles (Pollo and *Adh1*–1S /Coroico). The region of recombination and the parental sources is illustrated in Fig. 3.

It is evident from these data that particular alleles of the *Adh1* locus of maize represent a mosaic of different evolutionary histories. The estimation of gene phylogenies that depict the history of all *Adh1* alleles is by no means straightforward. A simple expedient is to eliminate all recombinant alleles from a phylogeny, although this expedient presents an inaccurate picture of the complexity of lineage evolution. The case of maize *Adh1* is not exceptional. Our studies of *Adh1* evolution in *Pennisetum glaucum* (pearl millet) reveal at least one recombinant allele in a sample of size 21 (Gaut & Clegg in press) and the classic studies of *Drosophila*

melanogaster by Krietman (1983) also revealed at least two recombinant alleles in a sample of size 11.

Despite the phylogenetic complexity of gene evolution, much can be learned from summary statistics like

θ, where $\hat{\theta} = sa^{-1}{}_{n-1}\, m^{-1}$ and $a^{-1}{}_{n-1} = \sum_{i=1}^{n-1} {}^{1}/_{i}$

where m is the number of sites in the sample and where s is the number of polymorphic sites. For neutral genes θ is an estimator of the important parameter $4N_e\,\mu$. Tests for neutrality at *Adh1* are designed to detect balancing selection or to detect a history of directional selection. Balancing selection at the *Adh1* locus can be detected by testing for heterogeneity of θ estimates along the *Adh1* sequence (e.g., see Hudson 1991). Gaut & Clegg (1993) tested for heterogeneity of θ across the *Adh1* gene by partitioning the gene into regions that were bounded by exon/intron junctions, and they found no evidence for heterogeneity.

Gaut & Clegg (1993) concluded that all regions of the *Adh1* gene were affected by a uniform evolutionary force, as would be expected if drift and mutation were the only forces governing *Adh1* evolution in maize. A selective sweep at the *Adh1* locus could result in uniform estimates of θ, despite a history of selection. In such a situation, the *Adh1* region would be expected to be depauperate in variation relative to other regions of the maize genome. Tests of heterogeneity in θ estimates among genes should reveal this pattern of selection. Estimates of θ from samples of several anonymous loci are available from maize (Shattuck-Eidens et al. 1990), and these were compared to the data from *Adh1*. The comparisons with other loci provide no evidence for a selective sweep at *Adh1*; however, it is important to note that these tests have low statistical power, and weak selection would go undetected (Gaut & Clegg 1993).

Because selection cannot be detected at the *Adh1* locus, we may tentatively regard θ as an estimate of the historical value of N_e for *Zea mays*. The statistic T_0 can also be estimated from θ, and this provides a picture of the depth of lineage diversity in the species. These estimates are summarised in Table 1 for both maize and pearl millet. The two species have rather different estimates of θ and N_e despite quite similar reproductive strategies and life histories. The *Adh1* sequence data provide useful insight into the evolutionary history of these two species, and they also reveal the relative importance of the various forces of evolution.

MONOCOT PHYLOGENY ESTIMATED FROM *rbc*L GENE SEQUENCE DATA

Over the past several years, a number of laboratories interested in plant molecular systematics have co-operated in the production of a large database for the estimation of seed plant phylogeny (reviewed in Clegg 1993). The focus of this effort has been the *rbc*L gene of the chloroplast genome that encodes the large subunit of the enzyme ribulose-1,5-bisphosphate carboxylase/oxygenase, responsible for the fixation of CO_2 in photosynthesis. The size of the *rbc*L gene is approximately 1431 bp (the length varies by several codons among seed plants at the extreme 3' end). In green algae and seed plants, the gene is not interrupted by introns, a fact that facilitates unambiguous sequence alignment.

Early studies of *rbc*L sequence evolution established that the *rbc*L gene evolves at a slow rate (Curtis & Clegg 1984), with quantitative estimates of synonymous rates of substitution for grass species at about 1.3×10^{-9} substitutions per nucleotide site per year (Zurawski et al. 1984). These estimates revealed the rate of *rbc*L gene evolution to be four- to five-fold slower than the average synonymous rate for plant or animal nuclear protein coding genes (Zurawski & Clegg 1987). Detailed statistical analyses of the rate and pattern of nucleotide substitution in *rbc*L, based on sequences sampled from cyanobacteria, algae, and flowering plants, established that this rate of evolution provided the appropriate window of resolution for the estimation of seed plant phylogeny (Ritland & Clegg 1987).

Two other developments combined to focus attention on the *rbc*L gene. The first development was the invention and rapid diffusion of PCR (polymerase chain reaction) technology, and the second development was the provision of a free kit of sequencing primers for the *rbc*L gene by Dr Gerard Zurawski (DNAX Research Institute, Palo Alto, California). Owing to the conservative rate of evolution, the primers could be almost universally

Table 1 Estimates of θ, N_e, and T_0 for anonymous loci (from Shattuck-Eidens et al. 1990) and for *Adh1* from *Zea mays* and estimates for *Pennisetum glaucum* for *Adh1*.

Species	θ	N_e	T_0
P. glaucum	0.0043	136 000	520 000
Z. mays	0.0210	660 000	1 900 000

employed for the PCR amplification and dideoxy DNA sequencing of flowering plant (angiosperm) *rbc*L genes. As a consequence of these developments, there has been a very rapid growth in the number of *rbc*L sequences available for analysis. By early 1992, more than 500 *rbc*L sequences had been determined from a range of seed plants, but with a primary concentration on the angiosperms. These data are currently being analysed and a full account is scheduled to be published in a special issue of the *Annals of the Missouri Botanical Garden* that will appear later in 1993 (Chase et al. in press).

Our laboratory has concentrated on the estimation of the phylogeny of the monocotyledons using a database of 104 sequences. Several lessons have become apparent in the course of analysing this relatively large dataset. The first lesson is that rates of molecular evolution for *rbc*L are heterogeneous over major monocot clades. Gaut et al. (1992) applied likelihood ratio tests, based on a maximum likelihood version of the relative rate test, to a 35-species subset of the monocot data and found that rates of molecular evolution varied over a more than five-fold range. The grass family has the highest rate of nucleotide substitution, while the palm family has the slowest rate. The rate of molecular evolution appears to be correlated with minimum generation time, so that long-lived species have slower rates than annual species. This poses a serious problem for phylogeny estimation algorithms that assume a molecular clock.

Initial experiments with parsimony algorithms, based on small subsets of the monocot data, revealed that the addition of a sequence sometimes induced large changes in estimated tree topology. This may have been a consequence of the long-edges attracting phenomenon (Felsenstein 1978; Hendy & Penny 1989). As more sequences were added so that all major clades had several sampled taxa, the estimated topology usually stabilised. The effect of increasing "taxon density" was most likely a consequence of resolving multiple substitutions at a site into single substitution events and providing better support for long branches as new sequences became associated with a long branch (Hendy & Penny 1989). Preliminary experiments using maximum likelihood (MLE) algorithms, applied to small datasets, suggest that MLE topologies do not exhibit the kind of instability seen for parsimony algorithms. We tentatively conclude that MLE algorithms are to be preferred when small sets of distantly related taxa are sampled.

Because the number of trees that can be produced for datasets of 100 or more species is incredibly large, there is no practical method for assuring that the global maximum solution has been identified for either parsimony or MLE algorithms. In fact, the computational complexity of MLE is so great as to preclude its use for datasets of this size. To illustrate this point, consider the work of Duvall et al. (1993) who submitted 79 sequences from the monocot dataset to a CRAY "Y-MP8/864" supercomputer using the program DNAML (PHYLIP 3.42; Felsenstein 1991). Fifty-one hours of CPU time were consumed before arriving at a MLE solution. To investigate the range of local solutions, it is essential to perform the analysis from a number of different initial conditions. To attempt to explore different local solutions, a 74-sequence subset of the dataset was submitted to the fastest existing supercomputer (Touchstone Delta Parallel Processing Supercomputer) using a recoded version of DNAML ("fastDNAML" version 1.03; Olson et al. 1992). Thirty-three different initial conditions (shufflings of the input order of the data) were explored, but 125 hours of CPU time were required to find the 33 local solutions. None of the 33 solutions were identical, based on MLE score and tree topology. Duvall et al. (1993) then added five more sequences to the maximum topology among the 33 solutions and used 50 random input orders of the data to find 79 species topologies. Six topologies were resolved and all six were found to be statistically equivalent based on the Kishino & Hasegawa test (1989).

The experiments of Duvall et al. (1993) illustrate the prohibitive costs associated with the application of MLE to large datasets. They also illustrate the fact that the likelihood surface may have islands of statistically equivalent MLE solutions. Parsimony algorithms are more practical for large datasets; however, the costs associated with exploring the surface of potential solutions to identify families of solutions approaching a global maximum is still large. One may also employ parsimony to determine a family of solutions and then calculate the MLE scores for these solutions to use in hypothesis testing via the Kishino & Hasegawa test. Combined use of parsimony and likelihood methods may be a desirable approach, because likelihood also provides estimates of branch lengths with their associated variances. The use of parsimony with bootstrapping on large datasets is too demanding of computer time to be practical. Thus, the use of likelihood to estimate statistical uncertainty and parsimony to search the surface of solutions may represent better use of computer resources.

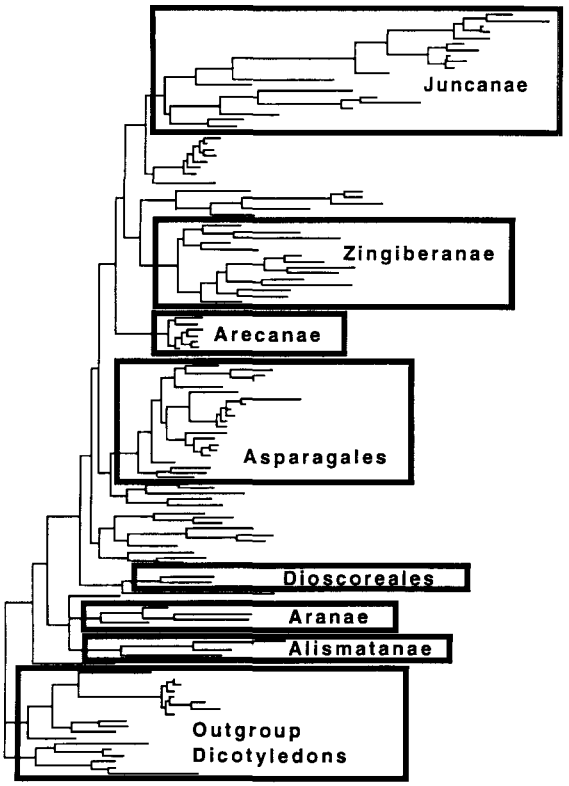

Fig. 4 Topology of one tree arbitrarily selected from a set of 109 equally parsimonious trees of 3932 steps (consistency index = 0.267; retention index = 0.633). The trees were produced by analysis of 1428 base pairs of DNA sequence data of *rbc*L sampled broadly across the monocotyledons (104 species) with 22 species of dicotyledons used as an outgroup. Seven major clades (orders and superorders) preserved as monophyletic groups in all 109 trees are boxed and identified.

Despite these complexities, the estimated trees contain much useful information about plant evolution. Figure 4, which presents one of 109 equally parsimonious trees estimated from the monocot *rbc*L dataset using PAUP (Swofford 1990) (104 monocot species plus 22 outgroup species), illustrates this point. This analysis identified *Acorus* as the primal extant monocot, and it revealed that major monocot clades emerged early during monocot evolution (Duvall et al. in press). Even when the uncertainties of tree estimation are taken into account, there seems little question that these data have provided an estimate of monocot phylogeny that has much greater precision than any previous estimates.

DISCUSSION AND CONCLUSIONS

Molecular data have provided a new window on plant evolution. These data yield more precise estimates of evolutionary history than have all previous data available to students of plant evolution, and they lend themselves to quantitative treatment. Molecular data also provide powerful insights into the relative importance of various evolutionary forces. Despite these enormous advantages, molecular data pose many unresolved problems in data analysis. As noted in Clegg (1993), plant evolutionists have suddenly been able to gather data at a much more rapid rate than the data can be analysed and assimilated. There is a great need for more efficient computer algorithms and for strategies of data analysis that mix various algorithms to achieve greater economy.

One mix of algorithms that should receive additional attention is the combination of MLE and parsimony algorithms. Our experience leads us to recommend that parsimony be used to search the surface of potential solutions. MLE can then be employed to discriminate among the set of parsimony solutions and to identify the subset of "best solutions" via the Kishino & Hasegawa test. Other mixed strategies, using, for example, neighbour-joining and MLE, should also be explored. The development of tree estimation algorithms that deal efficiently with large datasets must certainly be a priority for future theoretical and computational work in this field.

Despite the limitations of present algorithms, much has been learned from the monocot *rbc*L dataset. (1) The analyses revealed the previously unsuspected fact that *Acorus* is the primal extant monocot lineage (Duvall et al. 1993). (2) Substantial variation in rates of nucleotide substitution have been uncovered among major monocot lineages (Gaut et al. 1992), and these may confound some tree estimation algorithms. (3) The analyses reveal that most monocot orders and superorders emerged during a restricted period early in monocot evolution. (4) The *rbc*L-derived monocot phylogeny is in substantial agreement with traditional systematic treatments of the monocots (Duvall et al. in press).

What is it about molecular data, and especially DNA sequence data, that provides higher quality information? There are several potential answers to this question, but an important answer is that molecular sequence data provide a direct measure of identity-by-descent as opposed to identity-in-state.

Most population genetic theory is framed in terms of identity-by-descent transitions through generations. As a consequence, experimental population geneticists now have data that correspond to the quantities of theory. This has facilitated the rich development and application of coalescence theory in population genetics. The work with *Adh1*, cited above, represents a first attempt to exploit these opportunities in the study of plant population genetics. In the course of this effort, we have learned that intracistronic recombination is an important factor in plant gene evolution, just as it appears to be in animal gene evolution. We have been unable to detect evidence for selection at maize *Adh1*, despite the ability to integrate over evolutionary time, but this has facilitated the estimation of N_e, and it has revealed the temporal depth of allelic diversity in maize.

Phylogenetic trees estimated from single gene lineages do not always correspond to the complex patterns of reticulate evolution experienced by some plant species. Clearly, in the case of avocado, a gene tree provides little or no insight into the complex genealogical relationships among cultivars and varieties. To gain some useful insight into organismic history, it is necessary to average over many genes (genetic loci). The avocado work suggests that averaging, coupled with simple cluster analysis, can yield a good approximation to organismic history, although we still need to explore the limitations of this approach.

When long periods of evolutionary time are considered, as for the estimation of monocot phylogeny, it appears that single uniparentally transmitted genes can provide a good estimate of evolutionary history. This is not surprising, because we expect the temporal depth of reticulate evolution to be very shallow compared to that of, say, monocot phylogeny. As a consequence, we expect all gene phylogenies to converge to a very similar representation (within the limits of sampling and experimental error) at the temporal depth of monocot phylogeny. The particular transmission pathways (e.g., maternal, paternal, biparental) matter little at this depth. In light of these considerations, it seems that molecular biology has given students of plant evolution "the best of all worlds" in the sense that appropriate choices of molecular methods and analytical techniques are available at virtually all temporal levels.

ACKNOWLEDGMENTS

Support of the following grants is gratefully acknowledged: NIH grant GM 45144 (MTC); NSF grant BSR–9002321 (MRD); NIH GM 155528 (BSG); and grants from the California Avocado Commission (MTC). The DNA sequences discussed in this paper have been deposited with the Genbank database.

REFERENCES

Bergh, B. O.; Scora, R. W.; Storey, W. B. 1973: A comparison of the leaf terpines in *Persea* subgenus *Persea*. *Botanical gazette 134*: 130–134.

Chase, M. W.; Soltis, D. E.; Olmstead, R. G.; Morgan, D.; Les, D. H.; Mishler, B. D.; Duvall, M. R.; Price, R. A.; Hills, H. G.; Qui, Y.-L.; Kron, K. A.; Rettig, J. H.; Conti, E. L.; Palmer, J. D.; Manhart, J. R.; Kress, W. J.; Karoi, K. G.; Clark, W. D.; Gaut, B. S.; Jansen, R. K.; Kim, K.-J.; Wimpee, C. F.; Smith, J. F.; Furnier, G. R.; Strauss, S. H.; Xiang, Q.-y.; Plunkett, G. M.; Soltis, P. S.; Swensen, S.; Williams, S. E.; Gadek, P. A.; Quinn, C. J.; Eguiarte, L. E.; Golenberg, E.; Learn, G. H.; Graham, S. W.; Barrett, S. C. H.; Dayanandan, S.; Albert, V. A. in press: Phylogenetics of seed plants: An analysis of nucleotide sequences from the plastid gene *rbc*L. *Annals of the Missouri Botanical Garden 80*.

Clegg, M. T. 1993: Chloroplast DNA sequences and the study of plant evolution. *Proceedings of the National Academy of Sciences U.S.A. 90*: 363–367.

Curtis, S. E.; Clegg, M. T. 1984: Molecular evolution of chloroplast DNA sequences. *Molecular biology and evolution 1*: 291–301.

Dennis, E. S.; Gerlach, W. L.; Pryor, A. J.; Bennetzen, L.; Inglis, A.; Llewellyn, D.; Sachs, M. M.; Feri, R. J.; Peacock, W. J. 1984: Molecular analysis of the alcohol dehydrogenase (*Adh1*) gene of maize. *Nucleic acids research 12*: 3983–4000.

Duvall, M. R.; Learn, G. H.; Eguiarte, L. E.; Clegg, M. T. 1993: Phylogenetic analysis of *rbc*L sequences identifies *Acorus calamus* as the primal extant monocotyledon. *Proceedings of the National Academy of Sciences U.S.A. 90*: 4641–4644.

Duvall, M. R.; Clegg, M. T.; Chase, M. W.; Clard, W. D.; Kress, J. W.; Zimmer, E. A.; Hills, H. G.; Eguiarte, L. E.; Smith, J. F.; Gaut, B. S.; Learn, G. H. in press: Phylogenetic hypotheses for the monocotyledons constructed from *rbc*L sequence data. *Annals of the Missouri Botanical Garden 80*.

Felsenstein, J. 1978: Cases in which parsimony or compatibility methods can be positively misleading. *Systematic zoology 27*: 401–410.

Felsenstein, J. 1991: PHYLIP (Phylogeny inference package) Version 3.4. Seattle, University of Washington.

Furnier, G. R.; Cummings, M. P.; Clegg, M. T. 1990: Evolution of the avocados as revealed by DNA restriction fragment variation. *Journal of heredity 81*: 183–188.

Gaut, B. S; Clegg, M. T. 1991: Molecular evolution of alcohol dehydrogenase 1 in members of the grass family. *Proceedings of the National Academy of Sciences U.S.A. 88*: 2060–2064.

Gaut, B. S.; Clegg, M. T. 1993: Molecular evolution of the *Adh1* locus of the genus *Zea*. *Proceedings of the National Academy of Sciences U.S.A. 90*: 5095–5099.

Gaut, B. S.; Clegg, M. T. in press: Nucleotide polymorphism in the *Adh1* locus of pearl millet (*Pennisetum glaucum*) (Poaceae). *Genetics*.

Gaut, B. S.; Muse, S.; Clark, W. D.; Clegg; M. T. 1992: Relative rates of nucleotide substitution at the *rbc*L locus of monocotyledonous plants. *Journal of molecular evolution 35*: 292–303.

Hendy, M. D.; Penny, D. 1989: A framework for the quantitative study of evolutionary trees. *Systematic zoology 38*: 297–309.

Hudson, R. R. 1991: Gene genealogies and the coalescent process. *Oxford surveys in evolutionary biology 7*: 1–44.

Kishino, H.; Hasegawa, M. 1989: Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in the Hominoidea. *Journal of molecular evolution 29*: 170–179.

Kopp, L. E. 1966: A taxonomic revision of the genus *Persea* in the Western Hemisphere. *Memoirs of the New York Botanic Garden 14*: 1–117.

Krietman, M. 1983: Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. *Nature 304*: 412–417.

Olson, G.; Natsuda, H.; Hagstrom, R.; Overbeek, R. 1992: FastDNAML version 1.0.3. Argonne, Illinois, University of Illinois, Urbana and Argonne National Laboratory.

Osterman, J. C.; Dennis, E. S. 1989: Molecular analysis of the *Adh1*-C^m allele of maize. *Plant molecular biology 13*: 203–212.

Ritland, K.; Clegg, M. T. 1987: Evolutionary analysis of plant DNA sequences. *American naturalist 130*: S74-S100.

Sachs, M. M.; Dennis, E. S.; Gerlach, W. L.; Peacock, W. J. 1986: Two alleles of maize alcohol dehydrogenase 1 have 3' structural and poly(A) addition polymorphisms. *Genetics 113*: 449–467.

Shattuck-Eidens, D. M.; Bell, R. N.; Neuhausen, S. L.; Helentjaris, T. 1990: DNA sequence variation within maize and melon: Observations from polymerase chain reaction amplification and direct sequencing. *Genetics 126*: 207–217.

Smith, C. E. Jun. 1966: Archaeological evidence for selection of avocados. *Economic botany 20*: 169–175.

Smith, C. E. Jun. 1969: Additional notes on the pre-conquest avocados in Mexico. *Economic botany 23*: 135–140.

Sneath, P. H. A.; Sokal, R. R. 1973: Numerical taxonomy. San Francisco, W. H. Freeman. Pp. 230–234.

Swofford, D. 1990: Phylogenetic analysis using parsimony version 3.0s. Champaign, Illinois, Illinois Natural History Survey.

Zurawski, G.; Clegg, M. T. 1987: Evolution of higher-plant chloroplast DNA-encoded genes: implications for structure-function and phylogenetic studies. *Annual review of plant physiology 38*: 391–418.

Zurawski, G.; Clegg, M. T.; Brown, A. H. D. 1984: The nature of nucleotide sequence divergence between barley and maize chloroplast DNA. *Genetics 106*: 735–749.