Granule-Bound Starch Synthase: Structure, Function, and Phylogenetic Utility

Roberta J. Mason-Gamer,† Clifford F. Weil,* and Elizabeth A. Kellogg†*

*Department of Biological Sciences, University of Idaho; and †Department of Organismic and Evolutionary Biology, Harvard University

Interest in the use of low-copy nuclear genes for phylogenetic analyses of plants has grown rapidly, because highly repetitive genes such as those commonly used are limited in number. Furthermore, because low-copy genes are subject to different evolutionary processes than are plastid genes or highly repetitive nuclear markers, they provide a valuable source of independent phylogenetic evidence. The gene for granule-bound starch synthase (GBSSI or *waxy*) exists in a single copy in nearly all plants examined so far. Our study of GBSSI had three parts: (1) Amino acid sequences were compared across a broad taxonomic range, including grasses, four dicotyledons, and the microbial homologs of GBSSI. Inferred structural information was used to aid in the alignment of these very divergent sequences. The informed alignments highlight amino acids that are conserved across all sequences, and demonstrate that structural motifs can be highly conserved in spite of marked divergence in amino acid sequence. (2) Maximum-likelihood (ML) analyses were used to examine exon sequence evolution throughout grasses. Differences in probabilities among substitution types and marked among-site rate variation contributed to the observed pattern of variation. Of the parameters examined in our set of likelihood models, the inclusion of among-site rate variation following a gamma distribution caused the greatest improvement in likelihood score. (3) We performed cladistic parsimony analyses of GBSSI sequences throughout grasses, within tribes, and within genera to examine the phylogenetic utility of the gene. Introns provide useful information among very closely related species, but quickly become difficult to align among more divergent taxa. Exons are variable enough to provide extensive resolution within the family, but with low bootstrap support. The combined results of amino acid sequence comparisons, maximum-likelihood analyses, and phylogenetic studies underscore factors that might affect phylogenetic reconstruction. In this case, accommodation of the variable rate of evolution among sites might be the first step in maximizing the phylogenetic utility of GBSSI.

Introduction

Plant molecular systematists rely upon a relatively small number of molecular markers for phylogeny reconstruction, including, for example, the chloroplast DNA (cpDNA) genes *mat*K, *ndh*F, and *rbc*L, and, less commonly, cpDNA intergenic spacers. The nuclear genome is usually represented by sequences from the highly repetitive rDNA arrays—either the genes themselves, for studies of ancient divergence events, or spacer regions, for analyses at lower taxonomic levels. The justifications for choosing these markers are strong. First, because the chloroplast genome and the rDNA arrays occur in high copy numbers, they are technically relatively easy to work with, thus allowing systematists to amass data for large numbers of taxa. Second, individual systematists working with these markers are able to interpret their data within enormous databases of sequences that have been gathered by the systematic community. An additional advantage, specific to chloroplast markers, is that because the chloroplast genome is clonally inherited, problems associated with recombinant gene copies will not be encountered. The problem plant molecular systematists are now addressing is the limited number of markers available.

There are several reasons why low-copy nuclear genes have not been used as frequently as highly repetitive genes or chloroplast DNA genes, aside from technical considerations and the availability of comparable sequences from other data sets. First, low-copy nuclear genes are often poorly understood. Knowledge of gene copy number and the extent of concerted evolution among copies is critical for a reasonable within-individual sampling strategy. Second, there may not be enough sequences available across an appropriate taxonomic range to provide an initial estimate of phylogenetic utility. The development of primers is therefore more difficult, and sequencing genes from a large number of taxa is a potentially risky use of time and money.

However, plant systematists have been investigating various low- or single-copy nuclear markers for potential phylogenetic use. Most of those studied so far are members of small multigene families. For example, the gene for cytosolic phosphoglucose isomerase duplicated prior to the origin of the genus *Clarkia* (Onagraceae), and the two paralogs support identical relationships among the subgenera (Gottlieb and Ford 1996). The gene for alcohol dehydrogenase is duplicated in *Paeonia* (Paeoniaceae), but the two copies support significantly different relationships among nonhybrid peony species (Sang, Donoghue, and Zhang 1997). Members of the small family of phytochrome genes appear to be evolving independently (Mathews, Lavin, and Sharrock 1995), and a survey of the grass family showed that they can provide significant phylogenetic information (Mathews and Sharrock 1996). Members of the small family of defense-related *betv*1 homologs undergo concerted evolution such that gene copies from within plant families, and often those from within species, form distinct monophyletic groups and provide useful phylogenetic

Key words: starch synthase, Poaceae, Gramineae, *waxy*, plant molecular phylogenetics, likelihood models, secondary protein structure.

Address for correspondence and reprints: Roberta J. Mason-Gamer, Department of Biological Sciences, University of Idaho, Moscow, Idaho 83844-3051. E-mail: robie@uidaho.edu.

Mol. Biol. Evol. 15(12):1658–1673. 1998

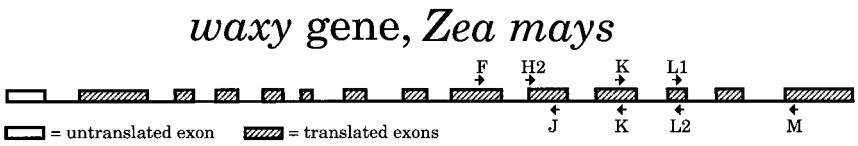

FIG. 1.—Schematic of the granule-bound starch synthase gene from *Zea mays.* Letters and arrows indicate designations, locations, and directions of the primers designed against four published grass sequences. The F and M primers mark the boundaries of the fragments used in the phylogenetic portion of this study.

information at the intrafamilial level (Wen et al. 1997). Sequence variation in the vicilin gene family has been examined in the legume family Fabaceae; phylogenetic analyses among some genera are complicated by the lack of homogenization among gene copies (de Miera and de la Vega 1998), while concerted evolution is extensive in other genera of this family (Doyle et al. 1986). Vicilin sequences have successfully been used for intergeneric phylogenetic analyses in *Theobroma* (Sterculiaceae); while copy number has not been fully characterized, multiple sequences from within species form monophyletic groups (B. Whitlock and D. Baum, personal communication). Clearly, evolutionary processes that affect the success or failure of phylogenetic analyses of plant nuclear loci vary among genes and taxa.

So far, fewer putatively single-copy genes have been examined for plants. We report on protein structure, nucleotide sequence evolution, and phylogenetic results based on a portion of the gene encoding granule-bound starch synthase (GBSSI; EC 2.4.1.11; fig. 1). The gene exists in a single copy in nearly all taxa in which it has been studied (e.g., Shure, Wessler, and Fedoroff 1983; Klosgen et al. 1986; Rohde, Becker, and Salamini 1988; Clark, Robertson, and Ainsworth 1991; van der Leij et al. 1991; Denyer and Smith 1992; Dry et al. 1992; Salehuzzaman, Jacobsen, and Visser 1993; Wang et al. 1995), although it appears to be duplicated in the Rosaceae (R. Evans, L. Alice, and C. Campbell, personal communication).

Many interrelated factors affect the phylogenetic utility of a gene, including, for example, the level of variation among taxa, the number of informative characters, the amount of homoplasy, and the presence of rate heterogeneity among taxa or sites. In turn, many of these factors are dependent on structural and functional constraints at the protein level. In this study, we examine GBSSI protein structure and nucleotide sequence evolution and conclude with a phylogenetic case study. First, we consider the protein structure of GBSSI along with the analogous microbial protein glycogen synthase (*glg*A; EC 2.4.1.11). Comparisons of the GBSSI and *glg*A amino acid sequences with the known three-dimensional structure of a mammalian protein, glycogen phosphorylase (EC 2.4.1.1), allow prediction of the nature and location of secondary structure. The identification of highly conserved inferred structural motifs influences the way in which sequences are aligned and allows the protein to be compared among widely divergent taxa. Second, we analyze the pattern of nucleotide sequence evolution using maximum-likelihood analyses based on a nested set of models of nucleotide substitution and among-site rate variation (e.g., Swofford et al. 1996; Frati et al. 1997; Sullivan, Markert, and Kilpatrick 1997). Because of position-dependent constraints on the protein sequence and the redundancy of the genetic code, the evolution of a protein-coding gene is expected to deviate from a simple model in which all nucleotide sites are equally free to vary. Furthermore, the physical characteristics of nucleotides lead to the expectation that some substitution types will occur more often than others. Likelihood models that account for unequal substitution frequencies and rate heterogeneity among sites are therefore expected to fit the data better than simpler models. The extent to which each parameter contributes to the improvement in fit allows identification of specific evolutionary processes that underlie the observed variation. Third, we perform cladistic parsimony analyses at several different taxonomic levels within the grass family Poaceae using a portion of the GBSSI gene. Exon sites are used for analysis of the family, while both exons and introns are used to estimate relationships among genera within tribes and among species within genera. These analyses demonstrate not only the potential utility, but also some possible pitfalls, of the use of protein-coding genes for phylogenetic studies.

Materials and Methods

Samples Within Grasses

We identified GBSSI (fig. 1) as a candidate for phylogeny reconstruction in the grasses because of evidence that it was single-copy and because seven sequences were available in GenBank, of which four were grasses. We used the grass sequences (wheat, barley, corn, and rice) to design primers, and generated 70 partial (1.3 kb) sequences of the gene from throughout Poaceae, sampling different taxonomic levels within the family (table 1). We included representatives of the two earliest diverging subfamilies (Anomochlooideae and Pharoideae), as well as all other major subfamilies except Chloridoideae. We also sampled intensively within the tribes Triticeae (including only diploid genera) and Andropogoneae and sampled multiple species within some genera.

DNA Sequencing

DNA was extracted following the method of Doyle and Doyle (1987) from between 1.5 and 2 g of fresh leaf material. Amplification reactions using the polymerase chain reaction were carried out with the F-for and M-bac primers (table 2 and fig. 1), which yield a product approximately 1.3 kb in length. In addition to the F-for and M-bac primers, six additional sequencing primers were designed for the F/M fragment (fig. 1 and

Table 1
Taxa Analyzed, Including Names, GenBank Accession Numbers, and Collection Numbers

Sample	GenBank	Collection
Panicoideae, Andropogoneae		
Andropogon gerardii	AF079235	PI 477973 (2)
Bothriochloa bladhii	AF079238	PI 301632 (1)
Capillipedium parviflorum	AF079239	PI 301780
Chrysopogon fulvus	AF079240	PI 199241 (2)
Chrysopogon gryllus	AF079241	PI 250984
Coix aquatica	AF079242	Kent s.n.
Cymbopogon flexuosus	AF079243	PI 209700
Cymbopogon jwarancusa	AF079244	PI 211159
Cymbopogon martinii	AF079245	PI 219582
Cymbopogon commutatus	AF079247	PI 180408
Cymbopogon popischilii	AF079248	PI 364476 (2)
Cymbopogon obtectus	AF079246	PI 257705
Cymbopogon refractus	AF079249	PI 301846
Cymbopogon schoenanthus	AF079250	PI 250644
Dichanthium aristatum	AF079252	PI 301994 (1)
Heteropogon contortus	AF079253	PI 364892 (1)
Hyparrhenia hirta	AF079254	PI 206889
Ischaemum santapaui	AF079255	PI 213265
Schizachyrium scoparium	AF079256	Kellogg V46
Sorghastrum nutans	AF079257	PI 315744 (1)
Sorghum bicolor	AF079258	PI 156549
Zea luxurians	AF079259	RS 95-12
Zea mays mays[a,b]	X03935	
Zea mays mexicana	AF079260	RS 95-10
Zea mays parviglumis	AF079261	RS 95-11
Panicoideae, Paniceae		
Pennisetum alopecuroides	AF079288	Park Seed 3650
Panicoideae, Arundinelleae		
Arundinella hirta	AF079236	PI 246756 (1)
Arundinella nepalensis	AF079237	PI 384059
Danthoniopsis dinteri	AF079251	PI 207548
Arundinoideae, Arundineae		
Hakonechloa macra	AF079292	Kellogg V21A
Pooideae, Triticeae		
Aegilops bicornis	AF079265	Morrison s.n.
Aegilops caudata	AF079262	G 758
Aegilops comosa	AF079263	G 602
Aegilops longissima	AF079266	Morrison s.n.
Aegilops searsii	AF079264	Morrison s.n.
Aegilops speltoides	AF079267	Morrison s.n.
Aegilops tauschii	AF079268	Morrison s.n.
Aegilops umbellulata	AF079269	Morrison s.n.
Aegilops uniaristata	AF079270	G 1297
Agropyron cristatum	AF079271	C-3-6-10(1)
Australopyrum retrofractum	AF079272	Crane 86146
Critesion californicum	AF079273	MA-138-1–40
Dasypyrum villosum	AF079274	PI 251478
Eremopyrum bonaepartis	AF079275	H5569(1)
Henrardia persica	AF079276	H5556
Heteranthelium piliferum	AF079277	PI 402352
Hordeum vulgare[a,c]	X07932	
Peridictyon sanctum	AF079278	KJ 248(1)
Psathyrostachys fragilis	AF079279	C-46-6-15
Psathyrostachys juncea	AF079280	PI 206684
Pseudoroegneria spicata	AF079281	PI 232117
Secale montanum	AF079282	PI 440654
Thinopyrum bessarabicum	AF079283	PI 531711
Thinopyrum elongatum	AF079284	PI 531719
Triticum aestivum[a,d]	X57233	
Triticum baeoticum	AF079285	Morrison s.n.
Triticum monococcum	AF079286	PI 2214134
Triticum urartu	AF079287	MIXED
Pooideae, Lygeae		
Lygeum sparteum	AF079289	Soreng 3698

Table 1
Continued

Sample	GenBank	Collection
Pooideae, Meliceae		
Glyceria grandis	AF079291	Davis and Soreng s.n.
Melica cupannii	AF079296	PI 383702
Bambusoideae, Bambuseae		
Chusquea exasperata	AF079293	LC&al. 1093
Chusquea oxylepis	AF079294	LC 1069
Bambusoideae, Oryzeae		
Oryza sativa[a,e]	X65183	
Bambusoideae, Parianeae		
Eremitis sp. nov.	AF079295	LC&WZ 1343
Pariana radiciflora	AF079297	LC&WZ 1344
Pharoideae		
Pharus lappulaceus	AF079298	LC 1329
Anomochlooideae		
Anomochloa marantoidea	AF079290	Clark 1299
Dicotyledons		
***Solanum tuberosum* (Solanaceae)**[a,f]	X58453	
***Ipomoea batata* (Convolvulaceae)**[a,g]	U44126	
***Pisum sativum* (Leguminoseae)**[a,h]	X88789	
***Manihot esculenta* (Euphorbiaceae)**[a,i]	X74160	
Prokaryotes		
Escherichia coli[j]	AE000419	
Haemophilus influenzae[k]	U32815	
Agrobacterium tumefaciens[l]	L24117	
Bacillus stearothermophilus[m]	D87026	
Bacillus subtilis[n]	Z99119	
***Synechocystis* sp.**[o]	D90915	
Methanococcus jannaschii[p]	67600	

NOTE.—Boldface indicates sequences used for predictions of secondary protein structure.
[a] cDNA sequences.
[b] Klosgen et al. (1986).
[c] Rohde, Becker, and Salamini (1988).
[d] Clark, Robertson, and Ainsworth (1991); not included in phylogenetic analysis.
[e] Wang et al. (1994).
[f] van der Leij et al. (1991).
[g] S.-J. Wang, K.-W. Yeh, C.-Y. Tsai (unpublished data).
[h] Dry et al. (1992).
[i] Salehuzzaman, Jacobsen, and Visser (1993).
[j] Blattner et al. (1997).
[k] Tatusov et al. (1996).
[l] Uttaro and Ugalde (1994).
[m] Takata et al. (1994).
[n] Kunst et al. (1997).
[o] Kaneko et al. (1996).
[p] Bult et al. (1996).

Table 2
List of Grass-Specific GBSSI Primers

Primer	Sequence	Used for
F-for	TGCGAGCTCGACAACATCATGCG	Amplification, sequencing
H2-for	GAGGCCAAGGCGCTGAACAAGG	Sequencing
J-bac	ACGTCGGGGCCCTTCTGCTC	Sequencing
K-bac	GCAGGGCTCGAAGCGGCTGG	Sequencing
K-for	CCAGCCGCTTCGAGCCCTG	Sequencing
L1-for	GCAAGACCGGGTTCCACATGG	Sequencing
L2-bac	CGCTGAGGCGGCCCATGTGG	Sequencing
M-bac	GGCGAGCGGCGCGATCCCTCGCC	Amplification, sequencing

NOTE.—Approximate primer locations indicated in figure 1.

table 2). All primers worked successfully for the entire family, except in a few scattered instances.

Amplification of the F/M fragment began with an initial denaturation at 94°C for 1 min, followed by five cycles of 94°C for 45 s, 65°C for 2 min, and 72°C for 1 min, then an additional 30 cycles of 94°C for 30 s, 65°C for 40 s, and 72°C for 40 s. Amplification was completed with a 20-min elongation step at 72°C to maximize A-tailing and increase efficiency of cloning into T-tailed vectors. Ten- or 20-μl amplification reactions were run, containing *Taq* DNA polymerase (Gibco/BRL; 0.05 U/μl), the included buffer (1×), $MgCl_2$ (1.5 mM), and primers (each 1 μM). Undiluted total DNA, which varied in concentration among samples, was used as a template at 0.5 μl/10 μl of reaction volume. The entire reaction volumes were run out on 1% agarose minigels. Bands were cut out and cleaned using Geneclean (Bio 101) and resuspended in 5–10 μl sterile dH_2O.

Cleaned PCR products were cloned into Invitrogen's pCR II or pCR 2.1 vectors, or into Promega's pGEM-T Easy vector. Ligation, transformation, and plating were carried out following the manufacturers' instructions, except that ligation and transformation reaction volumes were halved. Plasmid preparations were carried out following the mini alkaline-lysis/PEG precipitation protocol given in ABI User Bulletin 18 of October 1991. Plasmids were checked on 1% agarose gels for presence of inserts, and concentrations were estimated by visual comparison with bands containing known amounts of DNA.

Sequencing reactions were carried out with the ABI Prism dye terminator cycle sequencing kit with AmpliTaq DNA polymerase, FS. Reactions were run according to manufacturer's instructions, except that the reaction volumes were halved. Reactions were cleaned with the ethanol/sodium acetate protocol provided with the kit instructions. Cleaned sequencing reactions were run on an ABI 370A or ABI 377 automated sequencer following ABI recommendations. Between 90% and nearly 100% of each fragment was sequenced at least once in each direction using eight primers on the ABI 370 or six primers (K-bac and K-for excluded) on the ABI 377.

Sequence chromatograms were checked and edited by eye in ABI's SeqEd, version 1.03, or in Gene Codes Corporation's Sequencher, version 3.0. For each taxon, the overlapping fragments were joined by eye (in SeqEd) or automatically (in Sequencher). Sequences were exported as text and aligned using CLUSTAL V, and alignments were checked and adjusted by eye. For some phylogenetic analyses, introns were edited out of the aligned sequences by eye. Exon sequences were translated in MacClade 3.0 (Maddison and Maddison 1992).

Structural Analysis

Twenty-eight GBSSI amino acid sequences from grasses were added to four dicot GBSSI sequences and seven microbial *glg*A sequences from GenBank (boldface entries in table 1). Alignments were carried out initially using the Genetics Computer Group (University of Wisconsin) computer program ''pileup'' (Devereux, Haeberli, and Smithies 1984). Amino acid changes were judged to be conservative or nonconservative based on sources of information compiled by Ron C. Beavis and David Fenyö at the PROWL World Wide Web site (http://prowl.rockefeller.edu/). In particular, we visually compared observed changes to an idealized projection of the Dayhoff mutation odds matrix by multidimensional scaling (fig. 4*d* of Taylor 1986) and a Venn diagram of relationships among amino acids based on a variety of physicochemical properties (fig. 3*a* of Taylor 1986).

Amino acid sequences were analyzed for potential secondary structure and threaded onto known three-dimensional protein structures using the Phd program suite (Rost, Sander, and Schneider 1994) and the H3P2 program (Rice and Eisenberg 1997). Threading algorithms search among known three-dimensional protein structures and compare query sequences to this database for similar patterns (order and length) of secondary structure. Secondary structure predictions were noted on the alignments and the alignments were adjusted by eye for maximal correspondence of predicted structures. Maize transposon-induced mutations were obtained and characterized as previously described (Weil et al. 1992).

Maximum-Likelihood Analysis of Nucleotide Sequence Evolution

ML analyses were used to assess 16 models of sequence evolution, representing the 16 pairwise combinations between four models of nucleotide substitution and four models of among-site rate variation (e.g., Swofford et al. 1996; Frati et al. 1997; Sullivan, Markert, and Kilpatrick 1997). The four nucleotide substitution models tested were: (1) the Jukes-Cantor (JC; Jukes and Cantor 1969) model, which assumes that all nucleotide substitutions are equally probable and that nucleotides occur in equal frequencies; (2) the Kimura two-parameter (K2P; Kimura 1980) model, which incorporates rates of transitions versus transversions estimated using ML, and in which nucleotides are assumed occur in equal frequencies; (3) the Hasegawa-Kishino-Yano (HKY; Hasegawa, Kishino, and Yano 1985) model, in which transition and transversion rates are allowed to differ as above, and which incorporates observed average nucleotide frequencies; and (4) the general time-reversible (GTR; Yang 1994*a*) model, which incorporates observed average base frequencies and allows for rate variation among six substitution types (A↔C, A↔G, A↔T, C↔G, C↔T, and G↔T) following their estimation using ML. The four models of among-site rate variation tested were (1) no variation among sites; (2) some proportion of sites, estimated via ML, being allowed to be invariable (I), with no additional rate variation (Hasegawa, Kishino, and Yano 1985); (3) rate variation being allowed to follow a gamma distribution (Γ; Yang 1994*b*) with a shape parameter estimated using ML; and (4) some sites being allowed to be invariable, with the possibility of rate variation at the remaining sites following a gamma distribution (I + Γ; Gu, Fu, and Li 1995; Waddell and Penny 1996). All parameters

were estimated using PAUP* 4.0d63 on a set of 20 trees: 10 chosen randomly from the equally weighted parsimony analysis and 10 from the analysis in which codon positions were weighted 7:9:4. The choice of trees used in parameter estimations should not be critical, as long as they represent reasonable representations of the data and parameters do not fluctuate greatly from one tree to the next (Swofford et al. 1996 and references therein). Likelihood scores between models of sequence evolution were compared using a likelihood ratio test (Yang, Goldman, and Friday 1995). Dicotyledons were excluded from the ML analysis, because a test for stationarity of nucleotide frequencies across sequences (using PAUP* 4.0d63) showed that their inclusion led to a significant deviation from stationarity.

Phylogenetic Analysis at Family, Tribal, and Generic Levels

Pairwise K2P distances were calculated using PAUP* 4.0d54, with deletions treated as missing data. Phylogenetic analyses were carried out using cladistic parsimony methods. Tree length, consistency index (CI), retention index (RI), and rescaled consistency index (RCI) were computed with uninformative characters excluded. MacClade 3.0 was used for examining patterns of variation among codon positions and between exons and introns. PAUP* 4.0d53–54 was used for the entire grass data set, including four dicotyledons as outgroups, and PAUP 3.1.1 (Swofford 1993) was used for subsets of grass sequences.

For the full data set, introns were excluded; analyses were done on aligned exons using heuristic searches with TBR branch swapping. Bootstrap support was estimated with 10,000 bootstrap replicates and no branch swapping. Three weighting strategies, in addition to equally weighted sites, were used for the full grass data set. First, positions 1:2:3 were weighted 7:9:4, based roughly on the estimated number of changes for trees obtained in the unweighted analysis. Second, positions 1:2:3 were weighted 4.15:5.75:1.5, based on the minimum numbers of changes in first, second, and third positions. Third, all sites were proportionally reweighted once on a scale of 0 to 1, based on the maximum RCI for each character estimated on the trees from the unweighted analysis.

Tribes Triticeae and Andropogoneae, as well as intergeneric samples of *Cymbopogon* and *Triticum* + *Aegilops*, were analyzed both with and without introns, with all characters equally weighted. Within genera, effects of individual introns and exons were examined by analyzing each one separately and by sequentially excluding each one. Support within tribes and genera was estimated using 1,000 bootstrap replicates with a heuristic search and TBR branch swapping. Within genera, support was further assessed using the decay index (Bremer 1988; Donoghue et al. 1992).

Results

Protein Alignments that Consider Potential Protein Structures

The length and order of predicted protein structures were conserved in our sample of partial amino acid sequences from 28 grasses, 4 dicot plant species, 6 eubacterial species (including a cyanobacterium), and one archeon (fig. 2). The greatest variation occurred in predicted loops, as might be expected (Rost and Sander 1994). Adjustment of the sequence alignments to account for predicted secondary structures indicated where insertion of gaps led to better sequence alignments overall. The adjusted alignments also highlight individual amino acids that were conserved across all taxa despite their location in regions of low overall protein sequence identity. These were often at positions expected to be important for beginning or ending secondary structures. In contrast, the amino acid sequence conservation within those secondary structures was often low in comparisons between plant and bacterial sequences.

Threading each of the longer sequences onto known three-dimensional protein structures indicated that the best fit was to mammalian glycogen phosphorylase (Barford, Hu, and Johnson 1991; Rost 1995). The fit varied from sequence to sequence, with the best alignment score being 3.75. (For comparison, human muscle glycogen phosphorylase threaded against itself produces an alignment score of 4.11.) Protein-threading alignment scores greater than 3.5 identify proteins of similar sequence, structure, and function in most cases (Rost, Sander, and Schneider 1994). Aligning the sequences prior to threading strengthened the alignment score to 4.60. The region between glycogen phosphorylase residues 439 and 695 matched up with the GBSSI and *glg*A proteins particularly well; this region is shown in figure 2. Within this region, highly or completely conserved amino acids occurred in positions that correspond to the glycogen phosphorylase active site (fig. 3).

Maximum-Likelihood Analyses

Likelihood models that account for rate variation, either by allowing for some proportion of invariant sites (I) or by allowing for gamma-distributed rate variation among sites (Γ), resulted in the greatest increase in ML scores (fig. 4). Gamma-distributed rate variation was the single parameter that provided the most improvement (e.g., compare JC to JC + Γ in fig. 4). Transition/ transversion bias was low, with estimates ranging from 1.057 to 1.314 (table 3). There was, however, noticeable heterogeneity among substitution types when six types were considered, as in the case of the GTR models. In the GTR + I + Γ model, for example, estimated proportional differences among rates of reversible change between A↔C, A↔G, A↔T, C↔G, C↔T, and G↔T are 1.12, 2.71, 0.78, 1.57, 4.67, and 1.00, respectively (table 3). Allowing for differences in base frequencies also led to an improved model (e.g., compare the K2P and HKY models; fig. 4). Estimated frequencies of A, C, G, and T were 0.208, 0.296, 0.347, and 0.149, respectively.

The model that incorporates the most parameters, GTR + I + Γ, fit the data best, as expected (ML score = −8,340.281). The slightly simpler GTR + Γ model, lacking just one free parameter relative to GTR + I + Γ, had the second-best ML score (−8,349.931). Because of the nested relationship between the models, they can

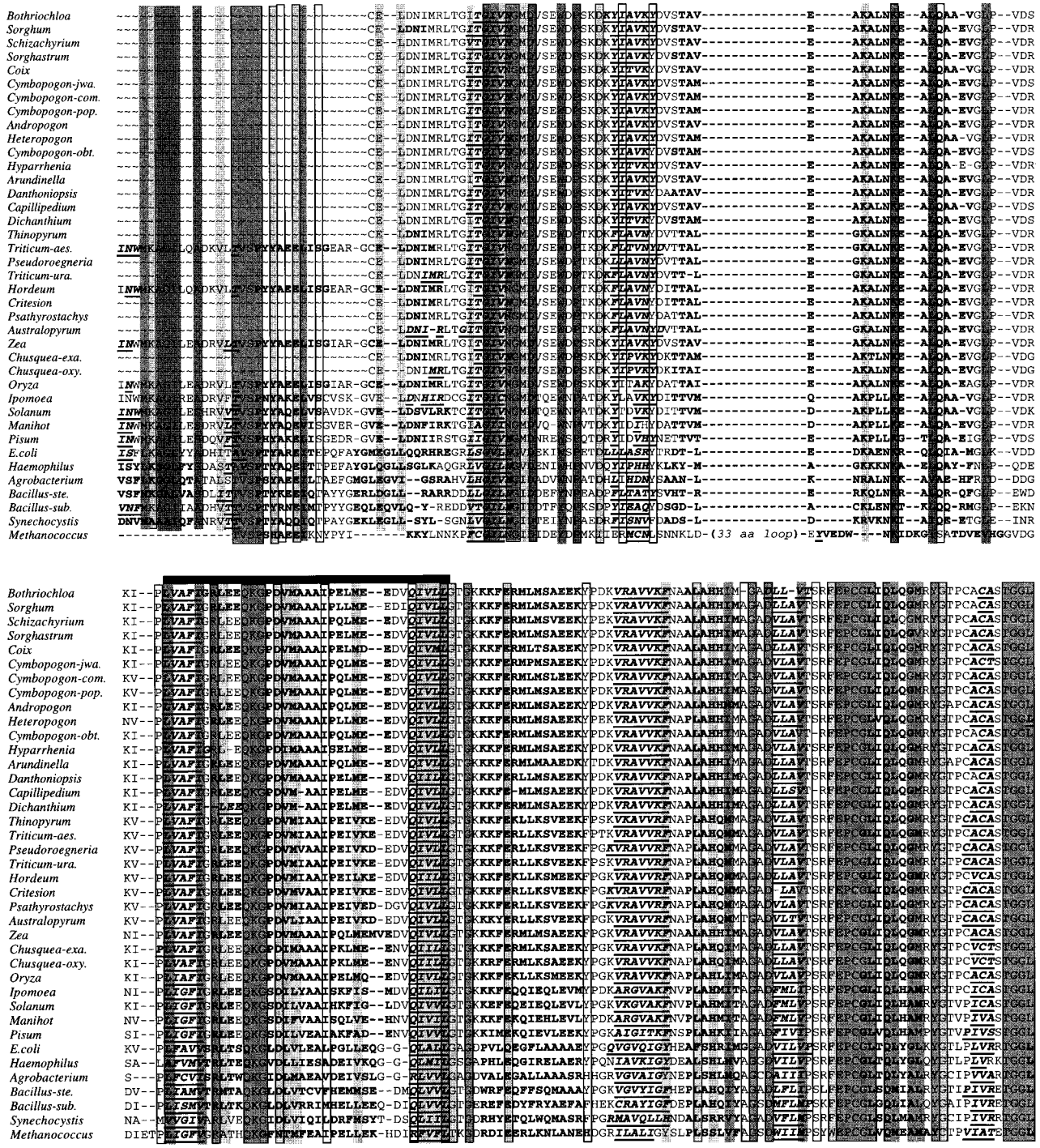

FIG. 2.—Partial alignment of starch and glycogen synthase proteins, showing predicted secondary structures. Alignment was carried out as described in *Materials and Methods.* Amino acids that are not included in the alignment (~) and sequence gaps (–) are indicated. Amino acids identical across all taxa are indicated by dark shading and a box; positions showing only conservative substitutions across all taxa are indicated by dark shading with no box; positions that are identical in all but one species are indicated by light shading and a box; positions that are conserved across all but one species are indicated by light shading and no box; and positions that are identical or conserved in all but two taxa are indicated by an open box. Predicted alpha helices are indicated in boldface; predicted beta structures are indicated in bold, underlined, italic type The black bar indicates the beta-alpha-beta structure referred to in the text. Taxa included in the alignment are shown in boldface in table 1.

be compared using a likelihood ratio test (Yang, Goldman, and Friday 1995). The observed difference in ML scores was used to obtain a test statistic, 2(9.65), which, when compared to a χ^2 distribution with one degree of freedom, indicated a significant difference ($P < 0.005$) between the two models. An even simpler model, such as HKY + I + Γ, includes four fewer free parameters than GTR + I + Γ, so a statistical comparison involves

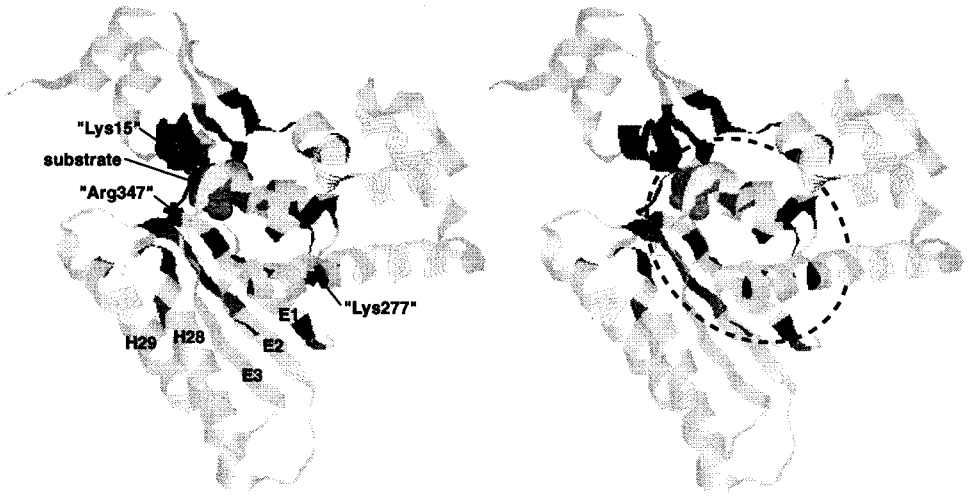

FIG. 3.—Partial stereo view of the mammalian glycogen phosphorylase three-dimensional structure, showing possible similarities to starch synthases. Regions shown as five parallel lines are absent in starch synthases. Positions indicated in darker gray correspond to either identical or conserved amino acids in the starch synthases of all plant and prokaryotic taxa examined. Residues indicated in medium gray are identical or conserved in all but one taxon (see fig. 2). E1–E3 indicate parallel beta structures; H28 and H29 are alpha helices. The beta-alpha-beta domain discussed in the text consists of E1, H28, and E2. ''Lys15,'' ''Lys277,'' and ''Arg374'' indicate the positions that these amino acids from the *Escherichia coli* protein would occupy The pocket that comprises the glycogen phosphorylase active site is indicated by the hatched oval.

more degrees of freedom. However, the difference in fit between the two models was again significant (test statistic = 2(26.502); $P < 0.005$).

Inclusion of dicotyledon GBSSI sequences led to the seemingly impossible result that the K2P model, which assumes equal nucleotide frequencies, always returned a better ML score than did the HKY model (result not shown), even though the HKY model incorporates the average observed nucleotide frequencies. This was because the addition of the dicotyledon sequences caused a significant deviation from stationarity of nucleotide frequencies across taxa ($P < 0.0001$), while grass sequences alone showed no such deviation ($P > 0.9999$). In this case, with the stationarity assumption violated, the incorporation of empirically derived nucleotide frequency values led to a poorer model than in which the frequencies of all four nucleotides were assumed to be equal.

Phylogenetic Analyses

Analysis of All Grasses

Less than 0.08% of the grass exon matrix (40 characters out of about 51,000) was coded as unknown. Across 773 aligned exon sites, there were a minimum of 606 changes at 395 variable sites, of which 274 were potentially informative. These included minima of 144

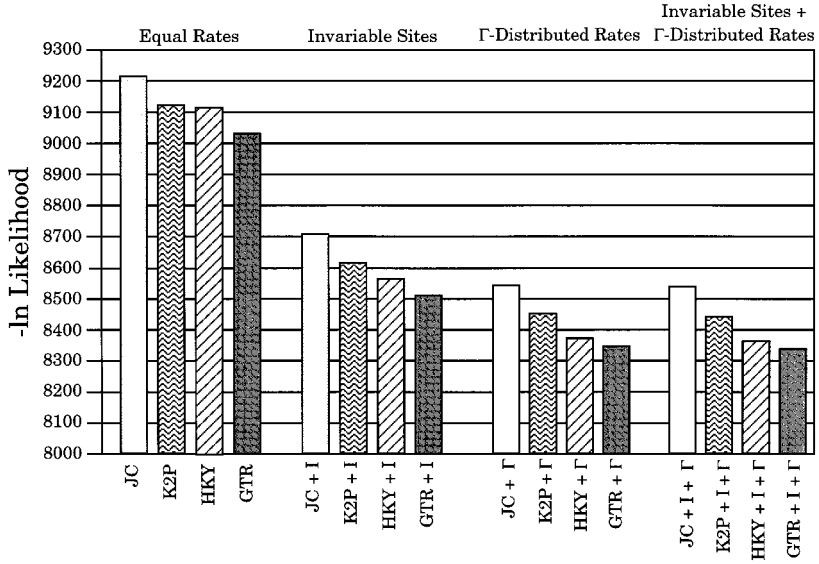

FIG. 4.—Histogram of the negative log-likelihood values obtained for 16 different models of nucleotide sequence variation. Lower values indicate a better fit between a model and the data. While each model parameter improves the likelihood score, as would be expected, the assumption that rates of change approximate a gamma distribution alone makes the greatest difference.

Table 3
Parameter Estimates for 12 Models of Sequence Evolution

Model	Negative Log-Likelihood	Gamma Shape Parameter (α)	Proportion of Invariable Sites	Ti/Tv Ratio[a]	R-matrix[b]
JC	9214.414	—	—	—	—
K2P	9123.419	—	—	1.057	—
HKY	9118.971	—	—	1.057	—
GTR	9031.347	—	—	—	1.49, 2.81, 0.59, 2.37, 5.14, 1.00
JC + I	8710.046	—	0.491	—	—
K2P + I	8617.713	—	0.491	1.083	—
HKY + I	8565.149	—	0.492	1.182	—
GTR + I	8513.441	—	0.490	—	1.01, 2.34, 0.62, 1.57, 4.34, 1.00
JC + Γ	8548.719	0.352	—	—	—
K2P + Γ	8454.457	0.350	—	1.107	—
HKY + Γ	8375.189	0.326	—	1.307	—
GTR + Γ	8349.931	0.342	—	—	1.13, 2.76, 0.77, 1.58, 4.55, 1.00
JC + I + Γ	8541.525	0.708	0.299	—	—
K2P + I + Γ	8447.034	0.708	0.300	1.110	—
HKY + I + Γ	8366.783	0.690	0.318	1.314	—
GTR + I + Γ	8340.281	0.757	0.327	—	1.12, 2.71, 0.78, 1.57, 4.67, 1.00

[a] Estimated transition-to-transversion ratio.
[b] Relative proportions of substitution types A↔C, A↔G, A↔T, C↔G, C↔T, and G↔T.

changes at the 103 variable first-codon-position sites, 90 changes at 78 second-position sites, and 372 changes at 204 third-position sites. Pairwise Kimura two-parameter corrected nucleotide distances within grasses ranged as high as 28.8%. At the amino acid level, including 257 aligned sites, there were a minimum of 228 changes at 131 sites, of which 81 were potentially informative.

A cladistic analysis of the exon sequences from all grass taxa, with first, second, and third codon positions weighted 7:9:4 respectively, resulted in 24 trees with a CI of 0.43, an RI of 0.75, and an RCI of 0.36. The strict-consensus tree (fig. 5) is well resolved, but bootstrap analysis shows little support for most of the tree. This was the case for all attempted weighting strategies. Nevertheless, the tree suggests relationships that are largely congruent with those from other data sets, including (1) monophyly of the Pooideae (Davis and Soreng 1993; Clark, Zhang, and Wendel 1995) and of the tribe Triticeae within it (Soreng, Davis, and Doyle 1990; Hsiao et al. 1995*a*); (2) the grouping of *Oryza* and the Bambusoideae with the Pooideae, supporting the ''BOP'' clade uncovered in other analyses (Clark, Zhang, and Wendel 1995); (3) monophyly of the Andropogoneae within the Panicoid subfamily (Davis and Soreng 1993; Clark, Zhang, and Wendel 1995); (4) monophyly of several lower-level clades, including *Triticum*/*Aegilops*, *Cymbopogon*, and *Zea*, along with species pairs representing *Arundinella*, *Chrysopogon*, and *Chusquea*; and (5) the basal positions of the Anomochlooideae and Pharoideae (Clark, Zhang, and Wendel 1995). (*Anomochloa* and *Pharus* are, however, on long branches [not shown], and their placement at the base of the tree is also consistent with the artifactual effect of long-branch attraction.) Different weighting strategies resulted in similar trees, except that in the unweighted analyses, *Chusquea* was placed near the base of the tree.

Analyses Within Tribes Andropogoneae and Triticeae

Levels of exon variation were similar for the two tribes examined. Among the 25 Andropogoneae sequences, there were minima of 45 changes at 41 variable first-position sites, 23 changes at 23 variable second-position sites, and 159 changes at 129 variable third-position sites. In all, 105 exon sites showed informative variation. A minimum of 65 changes occurred at 57 amino acid sites, of which 25 were informative. Among the 26 Triticeae sequences, there were minima of 46 changes at 40 variable first-position sites, 29 changes at 28 variable second-position sites, and 150 changes at 126 variable third-position sites; a total of 92 exon sites showed informative variation. As with the full data set, the GBSSI data within tribes provide well-resolved trees, but with low bootstrap support.

The gene tree for Andropogoneae (not shown) is similar to results obtained with the chloroplast gene *ndh*F (unpublished data). For example, in both GBSSI and *ndh*F, the tribe is monophyletic, with *Arundinella* as the sister taxon. Both genes find a clade comprised of *Bothriochloa*, *Capillipedium*, and *Dichanthium*, a group that is in agreement with extensive genetic, cytogenetic, and morphological work (deWet and Harlan 1970). Both genes suggest that the tribe resulted from a rapid radiation. The lack of support for relationships within the tribe (as in fig. 5) results not from extensive homoplasy, but rather from lack of substitutions, a result also found using *ndh*F sequences. Discrepancies between the GBSSI tree and the *ndh*F tree occur among branches with only one to three nucleotides changes.

Among the diploid genera of the Triticeae, it is harder to compare the GBSSI data with other data sets. While the diploid members of the tribe have been studied in numerous phylogenetic analyses (Hsiao et al.

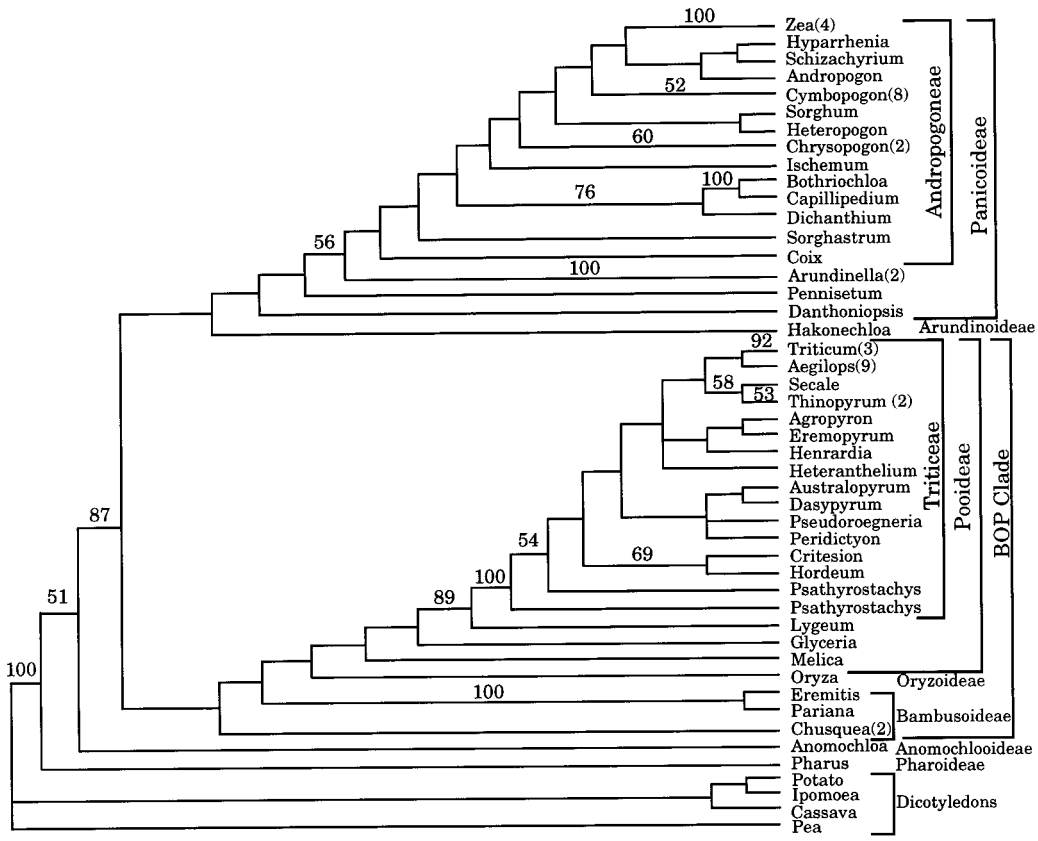

FIG. 5.—Results of phylogenetic analysis of exon sequences from the entire sample of grasses, with codon positions 1, 2, and 3 weighted 7:9:4; strict consensus of 24 trees. CI = 0.43, RI = 0.75, and RCI = 0.36. Four previously published dicotyledon sequences were used as outgroups. Bootstrap values higher than 50% are indicated above branches. Numbers in parentheses following some taxon names indicate cases in which multiple species were sampled per genus. Sampled subfamilies or tribes are indicated by brackets. The BOP clade represents a group (Bambusoideae, Oryzoideae, Pooideae) that has been detected in previous molecular analyses.

1995*b*; Kellogg and Appels 1995; Kellogg, Appels, and Mason-Gamer 1996; Mason-Gamer and Kellogg 1996*a*, 1996*b*; Petersen and Seberg 1997), all evidence to date points to different histories for different genomes or portions of genomes. The GBSSI tree is different from all published gene trees, but these are all significantly different from each other (Mason-Gamer and Kellogg 1996*b*). This suggests that GBSSI, which is on chromosome 7 in this tribe and is unlinked to any other gene studied to date, is tracking yet another portion of the Triticeae genome with its own history.

Among members of the Andropogoneae, introns cannot be aligned reliably, whereas introns among many Triticeae can. The notable exception within the Triticeae is intron 10. The first 5 and the last 18 bp of this intron can be unambiguously aligned for all taxa. Between these two regions, however, the sequences fall into two distinct classes, within which alignments are straightforward and between which alignments appear virtually meaningless. These two intron types divide the Triticeae into two groups of approximately equal size. The two groups delimited by the intron types do not match previously published phylogenetic hypotheses, nor are they in agreement with other characters within GBSSI. In this and in other, less extreme, cases of difficult alignment, we did not attempt to align the introns to one another, but instead aligned them as adjoining regions and coded the resulting gaps as missing data.

Analysis of Triticum/Aegilops

Twelve species from the very closely allied genera *Triticum* and *Aegilops* were analyzed both with and without intron characters included. Intron alignments were generally straightforward, but there were some regions of ambiguity due to insertions and deletions. With introns included, pairwise distances ranged from 0.81% to 8.42%. With introns excluded, distances ranged from 0.13% to 4.89%. Across all 12 species, 82 out of 771 exon sites were variable, with a minimum of 87 changes. Of these, 38 were potentially informative. First, second, and third codon positions included 18, 9, and 55 variable sites exhibiting minima of 19, 9, and 59 changes, respectively. Out of 587 aligned intron sites, 135 were variable; 76 of these showed potentially informative variation. Out of 257 amino acid sites, 25 were variable; 9 amino acid changes were potentially informative. Pairwise amino acid divergence ranged from 0.40% to 4.3%.

An unweighted cladistic parsimony analysis of all nucleotide positions yielded a single tree of length 210, with a CI of 0.61, an RI of 0.68, and an RCI of 0.51 (fig. 6). Analyses were rerun with exons excluded, with

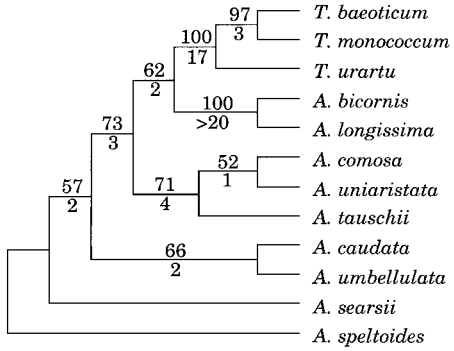

FIG. 6.—Results of cladistic parsimony analysis of a sample of *Triticum* and *Aegiops* species using equally weighted characters from exons and introns. Numbers above branches show bootstrap support; numbers below branches show decay indices.

introns excluded, with each individual exon and intron removed, and with each exon and intron alone. These analyses suggest that much of the phylogenetic signal reflected in the tree in figure 6 is provided by intron 12. In particular, the nodes grouping (1) *Aegilops bicornis* and *Aegilops longissima* with the three *Triticum* species; (2) *Aegilops comosa*, *Aegilops tauschii*, and *Aegilops uniaristata*; and (3) *Aegilops caudata*, *Aegilops umbellulata*, and *Aegilops searsii* with *Aegilops speltoides* at the base of the tree are unresolved when intron 12 is excluded from the analysis. All are present when intron 12 is analyzed alone.

Analysis of Cymbopogon

Eight species of *Cymbopogon* (tribe Andropogoneae) were analyzed with and without intron sites. Pairwise nucleotide divergence with introns included ranged from 2.25% to 11.13%. Pairwise divergence of exons alone ranged from 0.14% to 4.15%. The range of exon divergence is comparable to that seen in the *Triticum*/*Aegilops* group, but divergence when introns are included is somewhat higher. However, the introns within the genus are more difficult to align, and alignment ambiguity will artificially increase the estimated pairwise divergence values and the number of variable sites. Across the sample, 52 of 771 exon sites were variable; first, second, and third codon positions accounted for 9, 3, and 40 of the variable sites, and minima of 9, 3, and 43 changes, respectively. Twenty-one of the variable exon sites showed potentially informative variation. Out of 650 aligned intron sites, 146 were variable, with a minimum of 156 changes; 79 were potentially informative. There were a minimum of 13 amino acid changes at 13 sites; 4 of these were potentially phylogenetically informative. Pairwise amino acid divergence ranged from 0.39% to 3.15%.

Unweighted parsimony analysis of all characters resulted in a single shortest tree with a length of 261, a CI of 0.74, an RI of 0.74, and an RCI of 0.62 (fig. 7*A*). Analysis of introns alone yielded a single tree, identical to that from the full data set, with a length of 115, a CI of 0.75, an RI of 0.76, and an RCI of 0.57 (fig. 7*B*). Analysis of just exons yielded a single but different tree, with a length of 32, a CI of 0.79, an RI of 0.80, and an RCI of 0.63 (fig. 7*C*). Therefore, as is the case in the *Triticum*/*Aegilops* analyses, signal from the introns is driving the overall result. Unlike with *Triticum*/*Aegilops*, however, the source of the intron signal could not be narrowed down to a single intron. Exclusion of intron 10, 11, or 13 resulted in different tree topologies. Each of the introns analyzed alone also gave three different trees unlike that from the entire data set.

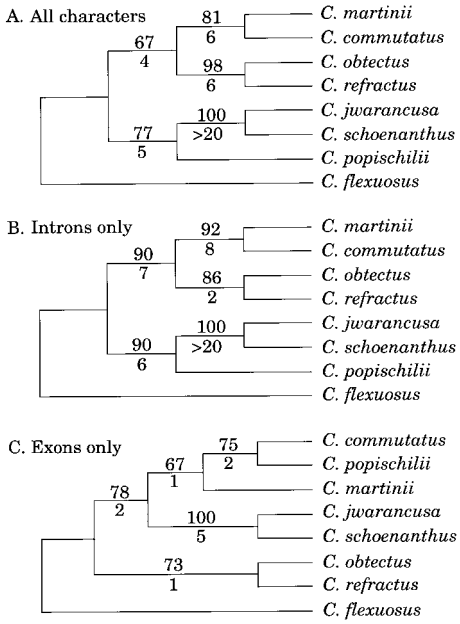

FIG. 7.—Results of cladistic parsimony analysis of eight *Cymbopogon* species with all characters equally weighted. The trees compare results obtained with (*A*) all characters, (*B*) only intron sites, and (*C*) only exon sites. Numbers above branches show bootstrap support; numbers below branches show decay indices.

Discussion

Analysis of Protein Structure

Amino acid sequence comparisons based on protein-threading alignments are theoretical and do not prove that conserved residues observed here are involved with the GBSSI or *glg*A active sites (Rost and Sander 1996); however, several additional lines of evidence support this idea. Biochemical studies of the *Escherichia coli glg*A protein have demonstrated that two lysine residues (positions 15 and 277 of the *E. coli* protein shown in fig. 3) are involved in the active site (Furukawa et al. 1990, 1994). Lys15 participates in substrate binding, and Lys277 participates in catalysis. These lysines thread onto positions at each end of the glycogen phosphorylase active site (fig. 3). The lysine analogous to *E. coli* Lys15 appears to be important in the plant proteins as well: a missense mutation, *waxy-C31*, that changes this lysine to an asparagine in the maize GBSSI protein, creates a null allele (unpublished data). Both the Phd and H3P2 threading algorithms identified strongest similarities between the GBSSI and *glg*A proteins and known proteins with beta-alpha-beta structures such as the one shown in figure 3. In addition, *Ac*/*Ds* transposon insertion alleles of the maize *waxy*

gene that affect what would be beta structures E1 and E2 (see fig. 3) revert rarely, while insertions that affect what would be analogous to alpha helix H28 revert frequently and to a variety of activity levels (Fedoroff, Wessler, and Shure 1983; Klosgen et al. 1986; Baran et al. 1992; Weil et al. 1992). *Ac*/*Ds* excisions that restore reading frame generally leave mutations adding one or two amino acids to a protein. These results suggest that structures like E1 and E2 may be more directly involved in GBSSI function than is H28. Furthermore, the different levels of activity seen for the transposon excisions from H28 may indirectly involve E1 and E2. Transposon excisions that make H28 longer should change the relative positions of E1 and E2 and also rotate the N-terminal end of the H28 helix, pulling the completely conserved and potentially important Arg347 residue (numbering as in the *E. coli* protein) away from the active site to varying degrees (see fig. 3).

Other studies have used information on molecular structure to infer alignment. For example, Bharathan et al. (1997) used structural characteristics to align a set of highly divergent homeobox genes. Similarly, the structure of 18S RNA has been used as an alignment guide for sequences representing a wide diversity of angiosperms (Soltis et al. 1997). Our study confirms the value of comparison of structural predictions for proteins of similar functions from distantly related organisms. Even though these predictions are somewhat speculative, they are still valuable for aligning coding regions that would otherwise be difficult to align with confidence because of low amino acid identity. It is reasonable to expect that distantly related proteins that catalyze the same reaction will have evolved appreciably in sequence while remaining similar in structure, with comparable active site nature and location, and conserved key amino acids. Evolution of sequences encoding protein secondary structures may represent an intermediate level of constraint between nonfunctional domains and enzymatically active sites. Proteins may change in sequence without losing function, but only in ways that do not disrupt critical secondary structures.

Once predicted structure has been used to help make alignments, comparative sequence data can be used to further infer sites that are functionally constrained. These may include crucial starting and ending positions for structures, as well as individual amino acids that are highly conserved between distant taxa despite being embedded in regions in which sequences have diverged. In an example of a similar approach, Kellogg and Juliano (1997) identified conserved regions of RuBisCO that may be functionally constrained, even though studies of the crystal structure do not indicate what the function might be. Regions of a protein that are under the least selective pressure are likely to be the most useful for comparisons of related species; this applies to both functional and structural constraints. For example, regions of a protein with little structure (such as a region with numerous loops) are more likely to accumulate changes because individual mutations are less likely to have far-reaching effects on how the rest of the protein folds.

Maximum-Likelihood Analyses of Nucleotide Sequence Evolution

Likelihood scores resulting from different models examined for the same tree are directly comparable, unlike, for example, tree lengths under different character-weighting schemes in cladistic parsimony analyses. Therefore, comparisons among ML models incorporating different parameters of sequence evolution can highlight evolutionary processes that contribute to the observed patterns of sequence diversity. In our GBSSI analyses, the most dramatic improvement in ML score accompanied the accommodation of among-site rate variation (I or Γ), with the greatest single improvement being achieved by allowing rate variation to follow a gamma distribution (Γ in fig. 4). This is consistent with the pattern of variation observed in the protein comparisons above; some amino acids (e.g., Lys15 and Lys277; fig. 3) are very strongly conserved and may not be free to vary at all, while others (e.g., those associated with loops) are under little constraint and are likely to have changed many times. Variation in the rate of evolution among sites is not surprising, given the functional constraints on proteins, and among-site rate variation has been observed for most coding sequences examined (e.g., Wakely 1994 and references therein). The need for assessing the magnitude of among-site rate variation for sequences used in phylogenetic analyses is clear. Numerous studies have shown that when rate variation is ignored, it can mislead many commonly used tree reconstruction methods (e.g., Jin and Nei 1990; Kuhner and Felsenstein 1994; Tateno, Takezaki, and Nei 1994; Yang, Goldman, and Friday 1994; Gaut and Lewis 1995; Sullivan, Holsinger, and Simon 1995; Swofford et al. 1996). Furthermore, the problem of long-branch attraction (Felsenstein 1978) can be exacerbated when rate heterogeneity is overlooked; this phenomenon has been demonstrated for a variety of tree-building methods (e.g. Tateno, Takezaki, and Nei 1994; Gaut and Lewis 1995; Huelsenbeck 1995).

Incorporation of unequal substitution probabilities or unequal base frequencies led to less dramatic (although still significant) improvements in ML score for GBSSI. The relative effects of different parameters of sequence evolution for this gene can be compared with those for other DNA sequences for which similar comparisons among ML models have been made. For example, the mitochondrial cytochrome oxidase gene of four collembolan insect families showed a result very similar to that for grass GBSSI sequences: the incorporation among-site rate variation following a Γ-distribution was the single change that caused the greatest improvement in ML score (fig. 4 in Frati et al. 1997). Among mitochondrial cytochrome *b* sequences from deer mice in the *Peromyscus aztecus* species group, on the other hand, incorporation of different probabilities for transitions and transversions led to the greatest improvement in ML score (fig. 4 in Sullivan, Markert, and Kilpatrick 1997).

The optimized parameters for the best ML model (last row in table 3) can be used to define the appropriate

Cymbopogon Intron 9

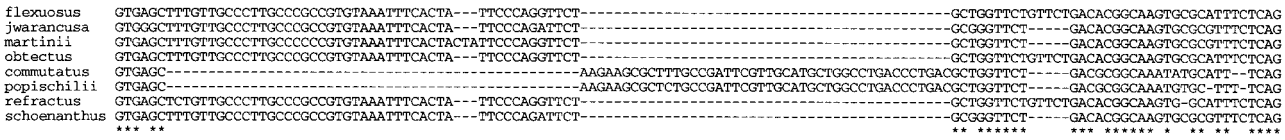

FIG. 8.—Alignment of intron 9 from eight *Cymbopogon* species. Asterisks indicate identical sites.

conditions for ML-based tree searches. For the grass GBSSI data set, not only did the most parameter-rich model (GTR + I + Γ) have the best score, as was expected, but its score was significantly better than those for simpler models, such as GTR + Γ and HKY + I + Γ. Therefore, the computational benefits of a simpler model would require a tradeoff of a significantly poorer fit of the model to the data. Many systematists have legitimate concerns about the computational intensity of ML tree searches, especially those involving many taxa and/or parameter-rich models. Even if ML-based tree searches are not performed, however, likelihood-based estimates of parameters of sequence evolution can highlight factors that may cause misleading results with other phylogenetic methods.

GBSSI in Molecular Phylogenetic Studies

We examined the utility of a single-copy gene for comparative evolutionary studies at several taxonomic levels. For phylogenetic reconstruction, the gene for granule-bound starch synthase illustrates several advantages and disadvantages, some probably common to many nuclear genes, and some that may be characteristic of this particular gene.

One potentially useful aspect of GBSSI is the high rate of change of nuclear gene introns. There is a clear need for nuclear markers for examining relationships among closely related species. The internal transcribed spacer (ITS) of the nuclear rDNA is often used for this purpose but is frequently not variable enough to distinguish species. ITS sequences in the Triticeae were identical for *A. speltoides* and *A. tauschii* (Hsiao et al. 1995*b*), while their GBSSI F/M fragments differed by 7.7%. The introns of the starch synthase gene not only exhibit higher levels of variation than the ITS region, but also potentially provide a greater number of characters as well. In this particular gene, there are 13 introns, usually ranging from 100 to 150 bp in length, while the two ITS regions total about 450 bp in the Triticeae (Hsiao et al. 1995*b*). GBSSI sequences have been successfully used to examine the closely related species of *Solanum* subsect. *Lycopersicum* (Peralta, Ballard, and Spooner 1997) and to determine relationships among members of the genus *Rubus* (L. Alice and C. Campbell, personal communication).

GBSSI is not appropriate for analyses at all phylogenetic levels. Introns diverge and become impossible to align long before exons have diverged enough to provide many variable characters. This may occur among species similar enough to be considered congeneric; alignment ambiguities exist within both *Cymbopogon* and *Aegilops.* Among more distantly related species, the exons of GBSSI can resolve relationships and produce phylogenies that are congruent with those from other genes (fig. 5). However, we have encountered a high level of among-site rate variation for grass exons, as well as a significant deviation from stationarity of nucleotide frequencies when dicot sequences are included. A reasonable phylogenetic analysis must account for such potentially misleading factors.

The intron-exon structure of GBSSI potentially allows it to be used for studies that span a range of taxonomic levels. However, if one is primarily interested in relationships between genera or families, sequencing through the many introns may not be an efficient way to generate data, and it may be more cost-effective to sequence either cDNA or another gene with fewer introns.

Recombination

For the alignment of intron 9 from *Cymbopogon* (fig. 8), *C. commutatus* and *C. popischilii* were very similar to one another but were difficult to align with the other taxa for a portion of the intron. Our solution was to separate the difficult-to-align portions and fill in the resulting gaps as missing data, but at the cost of losing the characters in this particular intron that distinguish these two taxa from the rest. The evidence from intron 9 suggesting that *C. commutatus* and *C. popischilii* are related is supported by three characters in the adjacent exon (10), and the analysis of exons alone places the two species together (fig. 7*C*). However, throughout the rest of the gene, there is just one other synapomorphy supporting the pair, and there are numerous characters (mainly in introns) supporting alternative placements of these two taxa. The species do not appear together in the shortest intron-only or intron+exon trees using the current alignment. If intron 9 is forced into an alignment with many mismatches, many characters then link *C. commutatus* and *C. popischilii,* but there is no evidence that the characters are homologous.

One possible explanation for the conflict between intron 9 and the rest of the gene is recombination, reflecting either a shared polymorphism or gene flow between the two species. Sometime after finding the curious pattern of intron variation, we discovered the following comment in the literature: ''Long racemes [in *C. popischilii*] tend to be associated with a large, but parallel-sided, lowermost pedicel, and may indicate introgression from *C. commutatus*'' (Clayton and Renvoize 1982, p. 765). Thus, gene flow between the two species had already been suggested based on morphological data; we have found additional evidence using molecular data.

The level at which introns are likely to be useful, namely among closely related species, is also the level at which recombinant genes are most likely to confound phylogenetic analysis. This is, in some ways, a potential disadvantage of nuclear markers, but recombination in fact provides a crucial window on biological history. Recombinant genes can be very difficult to detect, but when they are found, they may indicate ancient or contemporary polymorphisms, hybridization, or introgression.

Conclusions

Because protein function is often highly conserved, the nature and position of secondary structures can be used to aid in the alignment of amino acid sequences, even when amino acid identity among the sequences is very low. Strong constraints on protein structure and the resulting patterns of nucleotide diversity of coding sequences can potentially affect the outcome of phylogenetic analyses. However, when the processes underlying the observed patterns of nucleotide diversity can be examined and quantified, they can be accommodated such that they will be less likely to mislead phylogenetic results. In contrast to the conserved exons, introns are extremely variable, even among closely related species. They might provide phylogenetic information where other markers have failed, but they can, even within genera, become difficult to align.

Acknowledgments

The authors thank two anonymous reviewers for helpful comments on the manuscript. Thanks to Jack Sullivan for help with and discussion of maximum-likelihood analyses and Peter Lammers for helpful discussions. Lynn Clark and Jerrold Davis kindly provided several DNA samples. David Swofford allowed us to use, and to publish results from, PAUP* 4.0. This work was funded by NSF awards DEB-9419748 to E.A.K. and MCB-9506712 to C.F.W.

LITERATURE CITED

BARAN, G., C. ECHT, T. BUREAU, and S. WESSLER. 1992. Molecular analysis of the maize *wx-B3* allele indicates that precise excision of the transposable *Ac* element is rare. Genetics **130**:377–384.

BARFORD, D., S. HU, and L. JOHNSON. 1991. Structural mechanism for glycogen phosphorylase control by phosphorylation and AMP. J. Mol. Biol. **218**:233–260.

BHARATHAN, G., B.-J. JANSSEN, E. A. KELLOGG, and N. SINHA. 1997. Did homeodomain proteins duplicate before the origin of angiosperms, fungi, and metazoa? Proc. Natl. Acad. Sci. USA **94**:13749–13753.

BLATTNER, F. R., G. PLUNKETT III, C. A. BLOCHET et al. (17 co-authors). 1997. The complete genome sequence of *Escherichia coli* K-12. Science **277**:1453–1474.

BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution **42**:795–803.

BULT, C. J., O. WHITE, G. J. OLSENET et al. (41 co-authors). 1996. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. Science **273**:1058–1073.

CLARK, J. R., M. ROBERTSON, and C. C. AINSWORTH. 1991. Nucleotide sequence of a wheat (*Triticum aestivum* L.) cDNA clone encoding the waxy protein. Plant Mol. Biol. **16**:1099–1101.

CLARK, L. G., W. ZHANG, and J. F. WENDEL. 1995. A phylogeny of the grass family (Poaceae) based on *ndh*F sequence data. Syst. Bot. **20**:436–460.

CLAYTON, D. W., and S. A. RENVOIZE. 1982. Gramineae, part 3. Pp. 1–898 *in* R. M. POLHILL, ed. Flora of tropical East Africa. A. A. Balkema, Rotterdam.

DAVIS, J. I., and R. J. SORENG. 1993. Phylogenetic structure in the grass family (Poaceae) as inferred from chloroplast DNA restriction site variation. Am. J. Bot. **80**:1444–1454.

DE MIERA, L. E. S., and M. P. DE LA VEGA. 1998. A comparative study of vicilin genes in *Lens*: negative evidence of concerted evolution. Mol. Biol. Evol. **15**:303–311.

DENYER, K., and A. M. SMITH. 1992. The purification and characterisation of two forms of soluble starch synthase from developing pea embryos. Planta **186**:609–617.

DEVEREUX, J., P. HAEBERLI, and O. SMITHIES. 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12**:387–395.

DEWET, J. M. J., and J. R. HARLAN. 1970. Apomixis, polyploidy, and speciation in *Dichanthium*. Evolution **24**:270–277.

DONOGHUE, M. J., R. G. OLMSTEAD, J. F. SMITH, and J. D. PALMER. 1992. Phylogenetic relationships of Dipsacales based on *rbc*L sequences. Ann. Mo. Bot. Gard. **79**:249–265.

DOYLE, J. J., and J. L. DOYLE. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. **19**:11–15.

DOYLE, J. J., M. A. SCHULER, W. D. GODETTE, V. ZENGER, R. N. BEACHY, and J. L. SLIGHTOM. 1986. The glycosylated seed storage proteins of *Glycine max* and *Phaseolis vulgaris*. Structural homologies of genes and proteins. J. Biol. Chem. **261**:9228–9238.

DRY, I., A. SMITH, A. EDWARDS, M. BHATTACHARYYA, P. DUNN, and C. MARTIN. 1992. Characterization of cDNAs encoding two isoforms of granule-bound starch synthase which show differential expression in developing storage organs of pea and potato. Plant J. **2**:193–202.

FEDOROFF, N., S. WESSLER, and M. SHURE. 1983. Isolation of the transposable maize controlling elements *Ac* and *Ds*. Cell **35**:235–242.

FELSENSTEIN, J. 1978. Cases in which parsimony and compatibility methods will be positively misleading. Syst. Zool. **27**:401–410.

FRATI, F., C. SIMON, J. SULLIVAN, and D. L. SWOFFORD. 1997. Evolution of the mitochondrial cytochrome oxidase II gene in Collembola. J. Mol. Evol. **44**:145–158.

FURUKAWA, K., M. TAGAYA, M. INOYE, J. PREISS, and T. FUKUI. 1990. Identification of lysine 15 at the active site in *Escherichia coli* glycogen synthase. J. Biol. Chem. **265**:2086–2090.

FURUKAWA, K., M. TAGAYA, K. TANIZAWA, and T. FUKUI. 1994. Identification of Lys277 at the active site of *Escherichia coli* glycogen synthase. Application of affinity labeling combined with site-directed mutagenesis. J. Biol. Chem. **269**:868–874.

GAUT, B. S., and P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. Mol. Biol. Evol. **12**:152–162.

GOTTLIEB, L. D., and V. S. FORD. 1996. Phylogenetic relationships among the sections of *Clarkia* (Onagraceae) inferred

from the nucleotide sequences of *pgi*C. Syst. Bot. **21**:45–62.

GU, X., Y.-X. FU, and W.-H. LI. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. **12**:546–557.

HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating the human-ape split by a molecular clock of mitochondrial DNA. J. Mol. Evol. **22**:160–174.

HSIAO, C., N. J. CHATTERTON, K. H. ASAY, and K. B. JENSEN. 1995*a*. Molecular phylogeny of the Pooideae (Poaceae) based on nuclear rDNA (ITS) sequences. Theor. Appl. Genet. **90**:389–398.

———. 1995*b*. Phylogenetic relationships of the monogenomic species of the wheat tribe, Triticeae (Poaceae), inferred from nuclear rDNA (internal transcribed spacer) sequences. Genome **38**:211–223.

HUELSENBECK, J. P. 1995. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol. Biol. Evol. **12**:843–849.

JIN, L., and M. NEI. 1990. Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol. Biol. Evol. **7**:82–102.

JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 *in* N. H. MUNRO, ed. Mammalian protein metabolism. Academic Press, New York.

KANEKO, T., S. SATO, H. KOTANIET et al. (24 co-authors). 1996. Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions. DNA Res. **3**:109–136.

KELLOGG, E. A., and R. APPELS. 1995. Intraspecific and interspecific variation in 5S RNA genes are decoupled in diploid wheat relatives. Genetics **140**:325–343.

KELLOGG, E. A., R. APPELS, and R. J. MASON-GAMER. 1996. When gene trees tell different stories: the diploid genera of Triticeae (Gramineae). Syst. Bot. **21**:321–347.

KELLOGG, E. A., and N. D. JULIANO. 1997. The structure and function of RuBisCO and their implications for systematic studies. Am. J. Bot. **84**:413–428.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. **16**:111–120.

KLOSGEN, R. B., A. GIERL, Z. S. SCHWARZ-SOMMER, and H. SAEDLER. 1986. Molecular analysis of the waxy locus of *Zea mays*. Mol. Gen. Genet. **203**:237–244.

KUHNER, K. K., and J. FELSENSTEIN. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. **11**:459–468.

KUNST, F., N. OGASAWARA, I. MOSZER et al. (151 co-authors). 1997. The complete genome sequence of the gram-positive bacterium *Bacillus subtilis.* Nature **390**:249–256.

MADDISON, W. P., and D. R. MADDISON. 1992. MacClade 3.0. Sinauer, Sunderland, Mass.

MASON-GAMER, R. J., and E. A. KELLOGG. 1996*a*. Chloroplast DNA analysis of the monogenomic Triticeae: phylogenetic implications and genome-specific markers. Pp. 301–325 *in* P. P. JAUHAR, ed. Methods of genome analysis in plants. CRC Press, Boca Raton, Fla.

———. 1996*b*. Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). Syst. Biol. **45**:524–545.

MATHEWS, S., M. LAVIN, and R. A. SHARROCK. 1995. Evolution of the phytochrome gene family and its utility for phylogenetic analyses of plants. Ann. Mo. Bot. Gard. **82**:296–321.

MATHEWS, S., and R. A. SHARROCK. 1996. The phytochrome gene family in grasses (Poaceae): a phylogeny and evidence that grasses have a subset of the loci found in dicot angiosperms. Mol. Biol. Evol. **13**:1141–1150.

PERALTA, I. E., H. BALLARD JR., and D. M. SPOONER. 1997. ''Waxy'' gene intron phylogeny of tomatoes, *Solanum* subsect. *Lycopersicum* (Solanaceae) (Abstract). Am. J. Bot. **84**: 222.

PETERSEN, G., and O. SEBERG. 1997. Phylogenetic analysis of the Triticeae (Poaceae) based on *rpo*A sequence data. Mol. Phylogenet. Evol. **7**:217–230.

RICE, D., and D. EISENBERG. 1997. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. J. Mol. Biol. **267**:1026–1038.

ROHDE, W., D. BECKER, and F. SALAMINI. 1988. Structural analysis of the waxy locus from *Hordeum vulgare*. Nucleic Acids Res. **16**:7185–7186.

ROST, B. 1995. TOPITS: threading one-dimensional predictions into three-dimensional structures. Proc. Third Int. Conf. Intell. Syst. Mol. Biol. **3**:314–321.

ROST, B., and C. SANDER. 1994. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins **19**:55–72.

———. 1996. Bridging the protein sequence-structure gap by structure predictions. Annu. Rev. Biophys. Biomol. Struct. **25**:113–136.

ROST, B., C. SANDER, and R. SCHNEIDER. 1994. PhD—an automatic mail server for protein secondary structure prediction. Comput. Appl. Biosci. **10**:53–60.

SALEHUZZAMAN, S. N., E. JACOBSEN, and R. G. VISSER. 1993. Isolation and characterization of the cDNA encoding granule-bound starch synthase in cassava (*Manihot esculenta* Crantz) and its antisense expression in potato. Plant Mol. Biol. **23**:947–962.

SANG, T., M. J. DONOGHUE, and D. ZHANG. 1997. Evolution of alcohol dehydrogenase genes in peonies (*Paeonia*): phylogenetic relationships of putative nonhybrid species. Mol. Biol. Evol. **14**:994–1007.

SHURE, M., S. WESSLER, and N. FEDOROFF. 1983. Molecular identification and isolation of the *waxy* locus in maize. Cell **35**:225–233.

SOLTIS, D. E., P. S. SOLTIS, D. L. NICKRENT et al. (16 co-authors). 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Ann. Mo. Bot. Gard. **84**:1–49.

SORENG, R. J., J. I. DAVIS, and J. J. DOYLE. 1990. A phylogenetic analysis of chloroplast DNA restriction site variation in Poaceae subfam. Pooideae. Plant Syst. Evol. **172**: 83–97.

SULLIVAN, J., K. E. HOLSINGER, and C. SIMON. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA in sigmodontine rodents. Mol. Biol. Evol. **12**:988–1001.

SULLIVAN, J., J. A. MARKERT, and C. W. KILPATRICK. 1997. Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. Syst. Biol. **46**:426–440.

SWOFFORD, D. L. 1993. PAUP: phylogenetic analysis using parsimony. Version 3.1.1. Illinois Natural History Survey, Champaign.

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, and D. M. HILLIS. 1996. Phylogenetic inference. Pp. 407–514 *in* D. M. HILLIS, C. MORITZ, and B. K. MABLE, eds. Molecular systematics. 2nd edition. Sinauer, Sunderland, Mass.

TAKATA, H., T. TAKATA, T. KURIKI, S. OKADA, M. TAKAGI, and T. IMANAKA. 1994. Properties and active center of the thermostable branching enzyme from *Bacillus stearothermophilus.* Appl. Environ. Microbiol. **60**:3096–3104.

TATENO, Y., N. TAKEZAKI, and M. NEI. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. **11**:261–277.

TATUSOV, R. L., A. R. MUSHEGIAN, P. BORK, N. P. BROWN, W. S. HAYES, M. BORODOVSKY, K. E. RUDD, and E. V. KOONIN. 1996. Metabolism and evolution of *Haemophilus influenzae* deduced from a whole-genome comparison with *Escherichia coli*. Curr. Biol. **6**:279–291.

TAYLOR, W. R. 1986. The classification of amino acid conservation. J. Theor. Biol. **119**:205–218.

UTTARO, A. D., and R. A. UGALDE. 1994. A chromosomal cluster of genes encoding ADP-glucose synthase, glycogen synthase and phosphoglucomutase in *Agrobacterium tumefaciens*. Gene **150**:117–122.

VAN DER LEIJ, F. R., R. G. VISSER, A. S. PONSTEIN, E. JACOBSEN, and W. J. FEENSTRA. 1991. Sequence of the structural gene for granule-bound starch synthase of potato (*Solanum tuberosum* L.) and evidence for a single point deletion in the amf allele. Mol. Gen. Genet. **228**:240–248.

WADDELL, P. J., and D. PENNY. 1996. Evolutionary trees of apes and humans from DNA sequences. Pp. 53–73 *in* A. J. LOCKE and C. R. PETERS, eds. Handbook of symbolic evolution. Clarendon Press, Oxford.

WAKELEY, J. 1994. Substitution-rate variation among sites and the estimation of transition bias. Mol. Biol. Evol. **11**:436–442.

WANG, Z. Y., F. Q. ZHENG, J. P. GAO, D. P. SNUSTAD, M. G. LI, J. L. ZHANG, and M. M. HONG. 1995. The amylose content in rice endosperm is related to the post-transcriptional regulation of the *waxy* gene. Plant J. **7**:613–622.

WANG, Z. Y., F. Q. ZHENG, J. P. GAO, X. Q. WANG, M. WU, J. L. ZHANG, and M. M. HONG. 1994. Identification of two transposon-like elements in rice wx gene. Sci. China B. Chem. Life Sci. Earth Sci. **37**:437–447.

WEIL, C. F., S. MARILLONNET, B. BURR, and S. R. WESSLER. 1992. Changes in state of the *Wx-m5* allele of maize are due to intragenic transposition of *Ds*. Genetics **130**:175–185.

WEN, J., M. VANEK-KREBITZ, K. HOFFMAN-SOMMERGRUBER, O. SCHEINER, and H. BREITENEDER. 1997. The potential of *Betv1* homologues, a nuclear multigene family, as phylogenetic markers in flowering plants. Mol. Phylogenet. Evol. **8**:317–333.

YANG, Z. 1994*a*. Estimating the pattern of nucleotide substitution. J. Mol. Evol. **39**:105–111.

———. 1994*b*. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. **39**:306–314.

YANG, Z., N. GOLDMAN, and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. Mol. Biol. Evol. **11**:316–324.

———. 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. Syst. Biol. **44**:384–399.

BARBARA A. SCHAAL, reviewing editor

Accepted September 2, 1998