

Grass Evolution Inferred from Chromosomal Rearrangements and Geometrical and Statistical Features in RNA Structure

Gustavo Caetano-Anollés

Department of Crop Sciences, University of Illinois, Urbana–Champaign, IL 61801, USA

Received: 4 August 2004 / Accepted: 7 November 2004 [Reviewing Editor: Dr. David Pollock]

Abstract. The grasses (Poaceae) represent a monophyletic lineage that arose about 70 million years ago. The lineage contains about 10,000 species that differ widely in morphology and physiology. Species show striking differences in genome size, a feature important in the context of conservation of gene content and order (synteny and colinearity) and in the extension of genomic information directly from one grass species to another using comparative approaches. Grass diversification has been a contentious issue, as the exact branching order of the various subfamilies has been difficult to establish with standard methods. This motivated an evolutionary study of deep phylogenetic relationships based on the structure of coding and non-coding RNA molecules and on chromosomal rearrangements. Phylogenetic relationships in the grass family were inferred directly from the structure of RNA using cladistic principles and considerations in statistical mechanics. Coded attributes describing topological and thermodynamic information embedded in RNA molecules were treated as linearly ordered multi-state characters and were polarized by fixing the direction of character transformation toward molecular order. Intrinsically rooted phylogenies derived from the structure of signal recognition particle (SRP) RNA, the mRNA encoded by the early nodulation gene *enod40*, the small subunit of ribosomal RNA (rRNA), and the

internal transcribed spacer ITS1 of rRNA established an order for the diversification of major grass lineages, suggesting a sister relationship of the Pooideae and the PACCAD clade. This same conclusion was reached when large-scale chromosomal rearrangements derived from the comparative genetic mapping of cereal genomes were studied. Chromosomal complements aligned in the most parsimonious manner allowed identification and coding of characters depicting chromosomal translocations, insertions, and linkage block arrangements and the reconstruction of phylogenetic trees based on large-scale chromosomal structure. Congruent reconstruction of deep branching relationships using geometrical and statistical features of RNA structure and orthology and large scale chromosomal recombination events support assumptions of polarization in character argumentation, and fail to falsify the claim that extant grass chromosomes can be considered combinations of linkage blocks of an ancestor of the rice genome. Congruence also suggests that the universal tendency toward order in RNA and the search for the most parsimonious organization of be genome architecture appear to be mutually supported drivers of molecular evolution. The study clarifies the relationship of major clades in the grasses, shows that phylogenetic history can be reconstructed effectively from the combinatorial exchange of chromosomal linkage blocks, and reveals considerable phylogenetic signal embedded in the structure of signal polypeptide-coding mRNA molecules, describing an instance where mRNA structure is the subject of strong evolutionary constraint.

Key words: Chromosomal rearrangements — Cladistics — Grass evolution — Grass phylogeny — RNA structure

Introduction

The grass family (Poaceae) constitutes the fourth largest group of flowering plants. Grasses are spread throughout continents and habitats, are ecologically dominant, and cover about a fifth of the land on Earth. There are about 10,000 grass species that differ widely in morphology and physiology, including all the major cereals and most of the minor grains (Clayton and Renvoize 1986; Watson and Dallwitz 1992). Species also show striking differences in the size of their genomes, with DNA contents ranging from 0.5 to 40 pg DNA per 2C nucleus (Bennet and Leitch 2001). Besides polyploidization and duplication, genome size differences are often embodied in noncoding and repetitive DNA and result from mutational mechanisms of nucleic acid addition and loss, such as transposition (transposable element activity), spontaneous insertions and deletions, and chromosomal rearrangements (Petrov 2001). Genome size differences are important in the context of conservation of gene content and order (synteny and colinearity). Linkage mapping has shown that grass genomes can be described as combinations of linkage blocks corresponding (by orthology) to chromosomes of rice (*Oryza sativa* L), suggesting that it is possible to extend genomic information directly from one grass species to another using comparative genomic approaches (Devos and Gale 2000a). This has motivated evolutionary studies that trace changes in genome size throughout the history of diversification of the grass family (Bennetzen and Kellogg 1997; Kellogg 1998).

Classical and molecular phylogenetic analysis has shown that the grasses represent a monophyletic lineage that arose about 70 million years ago (Jacobs et al. 1999). Grass diversification, however, has been quite a contentious issue; the exact branching order of the various subfamilies has been difficult to establish (Kellogg 1998, 2001). A consortium of researchers recently combined molecular data from the plastid (chloroplast restriction sites and nucleotide sequences from the *ndhF*, *rbcl*; and *rpoC2* genes) and nuclear genomes (*phyB*, *waxy*, and ITS rRNA spacers) with morphological data (characters describing macromorphology, anatomy, biochemistry, and plastid restriction site variations) and produced a consensus phylogenetic tree (GPWG 2001). The tree was inferred from 2143 informative characters using parsimony, had bootstrap support (BS)

values of >90% for more than half of internal nodes, and represents the most comprehensive phylogenetic hypothesis of the grass family to date. It shows a basal placement of three subfamilies, the Anomochloideae, Pharoideae, and Puelioideae, one early major radiation (the BEP clade: bambusoids, ehrhartoids, and pooids) comprising the Bambusoideae, Ehrhartoideae, and Pooideae subfamilies, that is sister to a clade (PACCAD: panicoids, arundinoids, chloridoideae, centothecoids, aristidoids, and danthonioids) with all C4 lineages and several C3 plants that includes the Panicoideae and Chloridoideae subfamilies. While these results clarify the subfamily classification of the grasses, robust phylogenetic resolution at the base of the Poaceae has yet to emerge from molecular studies.

One difficulty with phylogenies is determining which characteristics are ancestral and which are derived. Defining directionality of character transformation is important for cladistics and generally requires the use of outgroups or external pre-existing hypotheses of relationship (Maddison et al. 1984). This enables tracing character change in the rooted trees, such as increases or decreases in genome size. However, phylogenetic inferences depend on how appropriate and closely related are outgroups and are affected by included taxa and reconstruction methods. An alternative is to polarize entire topologies by using directed characters. I have inferred phylogenetic relationships on the basis of shared and derived characteristics in RNA structure using cladistic principles and considerations in statistical mechanics (Caetano-Anollés 2001, 2002a,b). Molecules were characterized by attributes that describe the topology, stability, and uniqueness of folded conformations. These attributes were then treated as linearly ordered multi-state characters that were polarized by fixing the direction of evolutionary transformation toward molecular order. This approach unifies phylogenetics and structural biology, producing intrinsically rooted trees that “embed structure and function directly into phylogenetic analysis” (Pollock 2003). The approach allowed the reconstruction of a phylogeny of the living world (Caetano-Anollés 2002a) and the evolutionary tracing of RNA structure in ribosomes (Caetano-Anollés 2002b).

Phylogenetic analysis of nucleic acid structure involves three intimately related steps: (1) discovery and coding of characters, (2) establishment of an evolutionary direction of character transformation, and (3) reconstruction of phylogenetic trees, generally using maximum parsimony (Caetano-Anollés 2002b). These steps are illustrated in Fig. 1. RNA structures are first characterized using attributes that describe the overall “shape” (geometry) of molecules and “statistical” parameters that describe stability and mechanical features quantitatively. Shape attributes

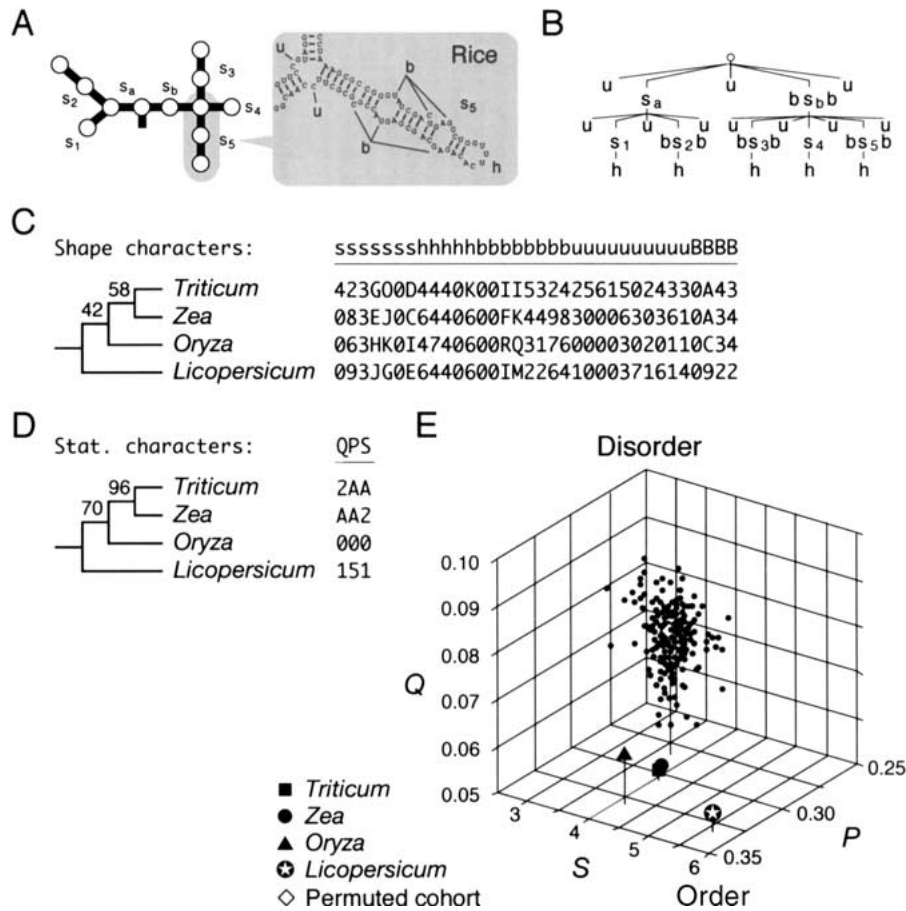


Fig. 1. Reconstructing phylogenies from structure: analysis of SRP RNA molecules. **A** Secondary structures from molecules of angiosperm origin were depicted as collections of loops (circles) connected by double-stranded stem tracts (lines). Every spatial component of secondary structure was characterized with shape attributes that describe their nucleotide length. Shape attributes describe double helical stems (*s*), hairpins (*h*), bulges and internal loops (*b*), unpaired segments (*u*), and the number of loops in coaxial stem tracts (**B**). **B** Structures were viewed as tree graph representations that map (5'-to-3') base pairs into internal nodes (*s*) and unpaired bases into leaves (*h*, *b*, and *u*). **C** Nodes and leaves define homologous "shape" characters that are coded, polarized by including a hypothetical ancestor, and arranged in a data matrix for cladistic analysis. Character coding is illustrated by the analysis of coaxial stem *s*₅ of rice (see **A**). The coaxial stem has a length *s* = 18 = 1 and contains (when visited in the 5'-to-3' direction) bulge loop segments of length *b* = 7, a hairpin loop of length *h* = 6, and bulge loops of length *b* = 6. Phylogenies were inferred using unconstrained maximum parsimony as the optimality criterion in

measure, for example, the nucleotide length of each and every spatial component of secondary structure, such as double helical stems and unpaired sequences, and the number of loops in coaxial stem tracts. Note that unpaired nucleotides can form unusual base-pairings or establish non-covalent interactions (Hermann and Patel 1999). These base pairs and interactions are involved in high-order three-dimensional motifs that are not considered in the structural models of this study. Statistical parameters include the Shannon entropy of the base-pairing probability matrix (*Q*), base-pairing propensity (*P*), and mean

PAUP. Characters were equally weighted except for stems that were weighted double to account for nucleotide number (*s*₆ was weighted 6 because PAUP accepts only 32 states per character). A single rooted most-parsimonious tree of 285 steps (CI = 0.789, RI = 0.355; *g*₁ = -0.073; PTP test, *p* = 0.001) was recovered following an exhaustive search. **D** Phylogenies were also reconstructed from statistical metrics that define quantitatively the stability and uniqueness of RNA secondary structure. These include the Shannon entropy of the base-pairing probability matrix (*Q*), base-pairing propensity (*P*), and mean length of helical stems (*S*). In this case, a single tree of 32 steps (CI = 0.937, RI = 0.875; *g*₁ = -1.210; PTP test, *p* = 0.001) was recovered after an exhaustive search. Bootstrap values (%) are shown for individual nodes. **E** A morphospace delimited by features *Q*, *P*, and *S* was used to confirm that evolution of SRP RNA increases conformational order over that inherently acquired by self-organization. Evolved sequences were more ordered than their corresponding permuted cohorts (50 replicates each) obtained by sequence randomization.

length of helical stems (*S*) (Fontana et al. 1993; Schultes et al. 1999; Ancel and Fontana 2000). *Q*, *P*, and *S* define a complete molecular morphospace, in which *Q* measures the number of conflicting inter- and intra-molecular interactions (a phenomenon known as *frustration*) during RNA folding, and *P* and *S* describe how extensively folded and ramified (multifurcated) are molecules (Schultes et al. 1999). *Q* is determined from the equilibrium partition function and portrays the thermodynamic likelihood of base pairing during the energy minimization process of folding. In phylogenetic analysis, attributes are con-

sidered “characters,” and the numerical values (and frequency distribution of values) they display “character states” (Page and Holmes 1998). To be useful, however, characters need to be appropriately coded so that they can provide maximum phylogenetic signal. Values were converted into a discrete alphanumeric format. For example, continuous valued states of statistical characters had to be gap-recoded into discrete data. Note that statistical characters describe the evolution of complete molecular ensembles and at present can only be used to trace spatial features individually in relatively simple molecules. Once coded, homologous characters (i.e., those sharing common ancestry) were compared. Structural characters used in this study transform from one state to another in linearly ordered and reversible pathways “polarized” by superimposing an evolutionary tendency toward structural order. This tendency is supported by phylogenetic congruence and considerations in statistical mechanics of molecules (see Materials and Methods) and can be interpreted as an evolutionary lock-in triggered by the branching of lineages in the trees (cladogenesis). This results in molecules that are less plastic but more modular. Polarization can be achieved by identifying the ancestral state or by using “step matrices” that assign transformation costs (measured in probabilities) to every possible change (Caetano-Anollés 2002b). Finally, hypotheses about character states and models of character evolution were transformed into hypotheses about evolutionary relationship of molecules using methods that minimize evolutionary change (maximum parsimony) (Page and Holmes 1998).

In this paper, I first describe the branching relationships of basal grass lineages with phylogenetic trees that were reconstructed directly from the structure of three functional RNA molecules for which a structural model is known and grass sequences are available: (1) *Signal recognition particle RNA (SRP RNA)*. These molecules represent small components (300–303 nucleotides in length) of cytosolic ribonucleoproteins capable of directing secretory proteins emerging from the ribosomes to proper cellular compartments (Keenan et al. 2001). A structural model of SRP RNA has been inferred from comparative sequence analysis and NMR spectrometry (Rosenblad et al. 2003). The model is highly consistent with structures predicted from aligned RNA sequences using free energy minimization and sequence covariation (Hofacker et al. 2002). (2) *mRNA molecules encoded by the early nodulation gene enod40*. *enod40* plays a central regulatory role in the *Rhizobium*–legume symbiosis; two small encoded peptides appear to be involved in the signaling process that underlies root nodule organogenesis (Sousa et al. 2001; Rohrig et al. 2002). Conserved secondary

structures with putative function were described in the *enod40* mRNA molecules (Hofacker et al. 2002), and enzymatic and chemical probing confirmed the predicted structures of three conserved domains believed to be regulatory (Girard et al. 2003). (3) *Small subunit ribosomal RNA (SSU rRNA)*. SSU rRNA is a fundamental functional component of ribosomes. Its structure matches models inferred by crystallography (Yusupov et al. 2001) and comparative sequence analysis (Wuyts et al. 2001).

In order to determine if the same branching pattern of grass lineages could be obtained using molecules lacking confirmed structure, I also studied the evolution of the internal transcribed spacer 1 (ITS1) of rRNA. A group of ITS1 sequences previously used to establish phylogenetic relationships in the grasses (Hsiao et al. 1999; GPWG 2001) and representative of major groups in Poaceae were selected. In the absence of a structural model inferred from comparative sequence analysis, minimum free energy structures (Zuker 1989; Mathews et al. 1999) were obtained using thermodynamic-based folding algorithms and used to infer phylogenetic relationships. These structural predictions should be relatively accurate, since ITS1 sequences are considered to be relatively short (Konings and Gutell 1995) and well evolved (Schultes et al. 1999).

Finally, I use evolutionary patterns inferred from the structure of these RNA molecules to evaluate phylogenetic trees generated using a completely different set of characters, this time describing major genomic rearrangements defined by comparative mapping. Using “circular” maps in which genomes were aligned and characterized in terms of ‘rice linkage blocks’ (Gale and Devos 1998a, b), I discover and code characters describing chromosomal inversions and translocations and instances of chromosomal orthology to more than one rice chromosome. The method is described in Fig. 2. These characters reflect large-scale syntenic relationships and were used to analyze how genomes representing the Pooideae and the Panicoideae were evolutionarily related to each other and to the small rice genome.

Materials and Methods

Sequences and Structures

The sequences and structures of SRP RNA and SSU rRNA were obtained from the SRP Database (<http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html>) or as DCSE alignments from the European rRNA Database (<http://www.psb.ugent.be/rRNA/index.html>), respectively. Sequences from internal transcribed spacers (ITS) of rRNA (Hsiao et al. 1999) were obtained from GenBank (accession nos. AF019849, AF019800, AF019807, AF019794, L11579, AF019788, AF019791, AF019818, AF019793, AF019787, AF019859, AF019829, AF019833, AF019842, AF019803,

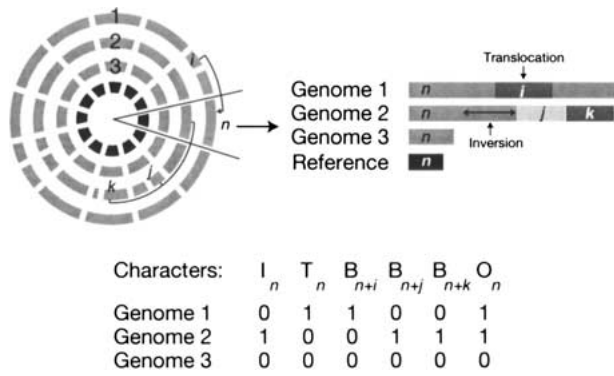


Fig. 2. Reconstructing phylogenetic trees from chromosomal rearrangements and distribution of rice linkage blocks in cereal genomes. The diagram shows a character-coding scheme, in which chromosomes (and corresponding linkage groups) related to a reference chromosome (n ; e.g., rice) (generally depicted as circular maps of genomes) are characterized by the absence (0) or presence (1) of inversions (I), translocations (T), and segments (“blocks”) related to other reference chromosomes (i, j, k) in their sequences (B). For each reference chromosome, the existence of heterologous blocks physically linked to homologous segments was also coded (O). In this paper, reference chromosomes are from rice.

AF0199854, and AF019817). The rice sequence was obtained from Takaiwa et al. (1995). The secondary structures of these sequences were predicted using the free-energy minimization method (Zuker 1989) with updated energy parameters (Mathews et al. 1999) using the Vienna package (Hofacker et al. 1994) or the *mfold* server (version 3.1; <http://www.bioinfo.rpi.edu/applications/mfold/old/rna/form1.cgi>).

Character Coding in RNA Structure. RNA structures are composed of stems (double-helical segments), loops (unpaired segments enclosed by stems), and external elements (unpaired segments that are not part of a loop) that are either free ends or joints that connect structural components of the molecules. Loops have different degrees (d) depending on the number of stems attached to them. Hairpin loops have $d=1$ and bulges and interior loops have $d=2$. All loops with $d>2$ are here called multi-loops and produced multifurcations in the molecules. Structures were here decomposed into structure elements (such as stems and unpaired segments) or sub-structural components (such as stem tracts and groups of bulges in a stem tract), attributes were characterized and coded as ordered multi-state characters using an alphanumeric format with numbers 0–9 and letters A–Q (to accommodate the number of maximum character states accepted by both phylogenetic analysis and character reconstruction programs), and matrices were encoded in the NEXUS format. Two general kinds of attributes were coded, those describing the detailed shape (herein defined as geometry) of molecules (*shape characters*), and those describing the stability and uniqueness of the folded molecules (*statistical characters*).

Shape characters used in this study measure the nucleotide length of each and every spatial component of secondary structure, such as stems and coaxial stem tracts (s), hairpin loops (h), bulges and internal loops (b), unpaired segments in multi-loops (m) or generic elements (u), and external segments (e), and the number of loops in a coaxial stem tract (B). In order to code shape characters, structures were coarse-grained into homeomorphically irreducible tree (HIT) graph representations (Fontana et al. 1993) that map in the 5'-to-3' direction base pairs into internal nodes and unpaired bases into leaves (Fig. 1). The nodes and leaves of these HIT graphs were used to code homologous shape characters (i.e. those sharing common ancestry) and construct data matrices for cladistic

analysis, with characters following the 5'-to-3' direction of the sequence. Substructural components such as coaxial stem tracts were generally characterized with more than one character (e.g., Fig. 1).

Statistical characters include statistics that provide global descriptions of elements of structure, such as the mean length of stems (S) and loops (L), the mean number of stems (N_s), the mean number of paired (N_p) and unpaired (N_u) nucleotides, the number of components (N_c), and the mean loop degree (D) (Fontana et al. 1993; Tacker et al. 1996). Note that the number of loops equals the number of stems, that $D=2-N_s/N_c$, and that the number of joints (N_j) equals N_c-1 . Statistical characters also include thermodynamic parameters such as the minimum free energy (G) and the minimum free energy of the thermodynamic ensemble (G_e). One set of especially useful statistical characters defines a complete structural morphospace, describing the Shannon entropy of the base-pairing probability matrix (Q), base-pairing propensity of individual structures (P) or the thermodynamic ensemble (P_e), and mean length of helical stems (S) (Huynen et al. 1997; Schultes et al. 1999). Q measures frustration during folding, and P and S describe how stable and multifurcated are molecules (Schultes et al. 1999). Q defines the uncertainty of base pairing and is here normalized to the length of the RNA sequence,

$$Q = -\frac{1}{Q_{\max}} \sum_{i=1}^{N-1} \sum_{j>i}^N p_{ij} \log_2 p_{ij} \quad (1)$$

where

$$Q_{\max} = \frac{1}{2} N \log_2 N \quad (2)$$

and p_{ij} is the probability of base-pairing between bases i and base j in the sequence that is calculated from the partition function (McCaskill 1990), and N is the length of the sequence in nucleotide residues (Schultes et al. 1999). For a perfectly defined structure $Q=0$ and for a folding with no preferred structure $Q=1$. P measures the number of base pairs normalized to sequence length and ranges from $P=0$ when there is no base pairing to $P=0.5$ when $N/2$ base pairs are formed. S measures the mean length of stems and its value ranges from $S=1$ for structures with single base pairs to $S=N/2$ for an idealized single hairpin with no loops.

The equilibrium partition functions and base pair probabilities were calculated using the Vienna package. Note that while RNA folding depends on the energetics of RNA molecules, the statistical mechanic parameters used to describe known structures are mostly dependent on molecular topology and are expected to be relatively insensitive to small adjustments in biochemically inferred energy parameters. This is not the case when molecules are modeled *de novo*, especially when sequences are not evolved.

Continuous-valued statistics were converted into linearly ordered multistate characters using the gap-recoding technique of Thiele (1993). Briefly, a rescaling function was used to rescore characters. This retains information on both rank order and size of gaps between character states. Raw data were ranked as ordered sets of states, standardized using an appropriate (usually square root) transformation if variances were unequal, and then range standardized using the maximum number of states allowed by the phylogenetic programs. This ensured that differences within and between characters were equalized with the transformation and range standardization procedure.

Comparison of Evolved Sequences with Randomized Derivatives. Random sequence cohorts with 50 replicates each were generated from the sequence of extant molecules by permutation using the heteropolymer randomization algorithms *stringgen* and *omrokgen* (available from V. Knudsen, USIT, University of Oslo). A Web interface to *omrokgen* is available at <http://nobai.uio.no/omrokgen/omrokgen.html>. Extant and randomized

derivatives were folded using the Vienna package and morphospace statistics were calculated for each individual sequence. An average value for each cohort and for the entire set was also calculated and plotted in a three-dimensional graph delimited by Q , P , and S statistics.

Characters Describing Large-Scale Chromosomal Rearrangements. Characters that describe non-redundant large-scale chromosomal rearrangements were identified in the consensus map that aligns 12 grass genomes in terms of rice linkage blocks (i.e., segments with markers that are colinear across other grass species) (Gale and Devos 1998a, b). For each genome, I studied the presence of translocations (T) and inversions (I), the physical connection of linkage blocks with each other (B), and the existence of chromosomal orthology (O) (see Fig. 2). There were five informative T characters describing known chromosomal translocations relative to rice (these involve rice chromosomes 3, 5, 6, 7, and 10) and three informative I characters describing the existence of inversions occurring in homologous sites. These characters were assigned states 0 and 1 for absence or presence of the rearrangement, respectively (and state 2 for characters in maize describing translocations involving segments orthologous to rice chromosomes 3 and 7 and accounting for the duplication of the maize genome). A total of 75 informative B characters were identified describing the absence (character state 0) or presence (state 1) of the same set of two linkage blocks in an individual chromosome. Finally, 12 informative O characters (corresponding to the 12 rice chromosomes) described the absence (state 0) or presence (state 1) for each rice chromosome of chromosomal segments orthologous to other rice chromosomes.

Character Argumentation for RNA Structure. Character evolution followed a model of character state transformation in which RNA structures with increased order were defined as being ancestral (plesiomorphic). Assumptions are supported by (1) congruence between trees generated from RNA sequence and those generated from structure (Caetano-Anollés 2001, 2002a, b), (2) the study of extant and randomized RNA sequences and the finding that molecular evolution enhances conformational order and diminishes frustration over that intrinsically acquired by self-organization (Stegger et al. 1984; Higgs 1993, 1995; Schultes et al. 1999; Steffens and Digby 1999; Gulyaev et al. 2002), (3) congruence between phylogenies generated from geomic and statistical characters (this work), (4) experimental verification of a molecular tendency toward order and stability using thermodynamic principles generalized to account for non-equilibrium conditions under a thermodynamical theory of evolution (Gladyshev and Ershov 1982), (5) correlation between thermal stability and the occurrence of structural motifs in natural nucleic acids (Kierzek et al. 2001), and (6) a large body of theoretical evidence that maps the structural repertoire of evolving RNA sequences from energetic and kinetic perspectives (reviewed in Fontana 2002), with some important predictions recently confirmed experimentally (Schultes and Bartel 2000). Hypothetical ancestral molecules were chosen as those having maximum base pairing, order, and thermodynamic stability (larger s , S , P , N_b , G , and G_c) and minimum frustration and destabilization from unpaired and unordered regions (Q , B , L , N_s , N_c , D , h , b , u , m , and e). Character attributes represent transformation pathways and hypotheses of relationship that are falsifiable and link character states to each other using basic evolutionary assumptions or axioms (Bryant 1991). Phylogenetic analysis of RNA structure rests on a very simple (perhaps simplistic) model and on the auxiliary assumption that there is an evolutionary tendency towards order (hypothesis of polarization). This tendency may represent an accurate depiction of generalized trends, but the model may fail to explain exceptions and departures to the trend. Shape phylogenies have shown discordance, such as the incorrect

placement of highly simplified rRNA structures from intracellular parasites (e.g., *Encephalitozoon cuniculi*) (Caetano-Anollés 2002b), and this is suggestive of a need for more refined evolutionary models that may account for other factors that impact structure.

Multi-state characters transform in this study from one character state to another in linearly ordered and reversible pathways. Order implies a distance relationship between character states in which costs related to the transformation of two non-neighboring states is larger than one step. In fact, costs are initially simply the difference between the integers that describe the states. However, polarization identifies the ancestral states in the character transformation series and this results in characters that are directional and show asymmetry between gains and losses. The choice of linearly ordered characters is appropriate for both shape and statistical RNA features. RNA structures normally change in discrete manner by nucleotide addition or removal and this results in one-step extension or contraction of topological features. Insertion or deletions are possible but have a higher cost associated with them. They can remove entire substructures and are generally rare events. In this study, we consider that the cost of insertions and deletions is proportional to their length. Statistical characters represent continuous-valued features, but are gap-recoded into discrete entities. Again, only linearly ordered characters describe transformations appropriately. Note that geometrical and statistical features cannot be described appropriately using unordered characters.

Step matrices were also used to assign transformation costs to changes between character states (Caetano-Anollés 2002b). Patterns of character change were inferred with the "state changes and stasis" feature in MACCLADE v. 3.08 (Maddison and Maddison 1999), and relative frequencies were plotted in bubble diagrams and then converted to a transformation type using functions described by Wheeler (1990).

Characters were equally weighted. However, when considering shape characters, stems were weighted double compared to unpaired segments to account for nucleotide number. Note that a more realistic weighting scheme should also consider evolutionary rates of change in segment length, but this requires the measurement of evolutionary parameters along individual branches of the tree and the development of an appropriate quantitative model.

Character Argumentation for Large-Scale Chromosomal Rearrangements. Binary and three-state linearly ordered characters describing large-scale chromosomal rearrangements were first analyzed without invoking a hypothesis of polarization. This generated unrooted trees. In subsequent analyses, characters were polarized in one or the other direction, generating competing rooting hypotheses. In one case, linkage blocks present in rice were considered ancestral and re-arrangements that result in translocations, insertions, and re-distribution of rice linkage blocks were considered derived events. This polarization scheme supports the claim of Gale and Devos (1998b) that linkage blocks "free-standing" in rice can be found in different chromosomal environments, i.e., that extant chromosomes analyzed are combinations of linkage blocks of an ancestor of the rice genome. The polarization scheme in the opposite direction, however, supports other equally probable scenarios. Rooting hypotheses were then tested by phylogenetic reconstruction and by comparing tree lengths and topologies.

Tree Search and Analysis. The evolutionary history of the structure of RNA and large-scale chromosomal rearrangements were reconstructed using maximum parsimony methods in PAUP* v. 4.0 (Swofford 1998). The ANCESTATES command was invoked to define ancestral character states and polarity of character transformation. When exhaustive and branch-and-bound searches were not possible, phylogenetic trees were derived from heuristic searches using tree-bisection-reconnection (TBR) branch swapping

and at least 10 replicates of random addition sequence. Phylogenetic reliability was tested by the nonparametric bootstrap method (Felsenstein 1985) implemented using $1-2 \times 10^3$ pseudoreplicates and by double decay (DD) analysis (Wilkinson et al. 2000) using the RADCON program (Thorley and Page 2000). DD-derived reduced cladistic consensus (RCC) support trees were used to describe the best-supported phylogenetic hypotheses. Partition tables are not shown. The structure of phylogenetic signal in the data was tested by the skewness (g), of the length distribution of 10^4 random trees and permutation tail probability (PTP) tests of cladistic covariation using 10^3 replicates. The homogeneity of partitions was analyzed using a modified Mitchenich-Farris index of incongruence among data sets and 10^3 heuristic search replicates (Farris et al. 1995).

Results

Phylogenetic Analysis of SRP RNA Structure

Rooted phylogenies describing the evolution of SRP RNA structure were generated using shape and statistical characters from structures representing three grass species and tomato (*Lycopersicon esculentum* L.) (Fig. 1). The number of taxa included in the analysis was limited by relevant sequences deposited in the SRP Database. Distribution of cladogram lengths ($p < 0.01$) and PTP tests ($p = 0.001$) indicate that there was significant phylogenetic signal in structural data. Trees derived from shape and statistical characters had congruent topologies, but branches in the latter were better supported. The trees showed that SRP RNA from rice diverged earlier than molecules from wheat (*Triticum aestivum* L.) and maize (*Zea mays* L.). The model of character transformation in this system was tested using the molecular morphospace defined by Schultes et al. (1999) (Fig. 1). Each SRP RNA sequence was used to generate a cohort of 50 permuted sequences, and the structure of these randomized sequences were compared with the original structure. Structures from evolved SRP RNA sequences had lower Q and larger P and S values ($p = 0.05$) than randomized derivatives generated by permutation using heteropolymer randomization algorithms. This observation confirmed that evolved structures were more ordered than those expected from randomized sequences.

Phylogenetic Analysis of Conserved Motifs in *enod40* mRNA

Phylogenies were also reconstructed from the structure of *enod40* mRNA molecules (Fig. 3). These structures were analyzed in five grass species and the legumes *Glycine max* (L.) Merr. (soybean) and *Medicago truncatula* (L.) Gaertn. (barrel medic) (Table 1). There was considerable phylogenetic signal in the data (cladogram length distribution, $p < 0.01$; PTP tests, $p = 0.001$). Phylogenies were congruently

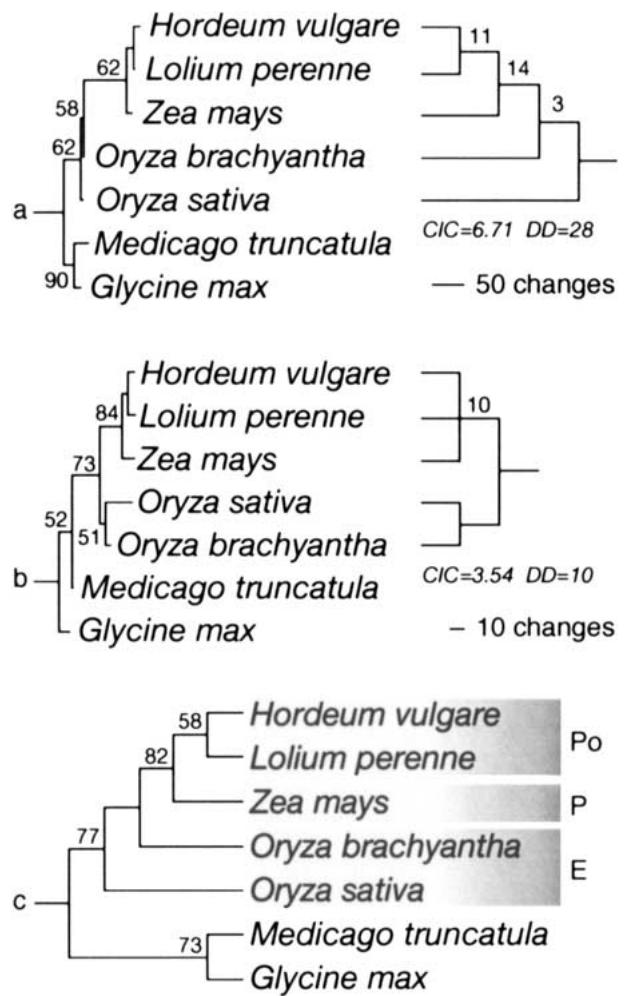


Fig. 3. Reconstructing phylogenetic trees from the structure of *enod40* mRNA molecules. Trees were reconstructed from 15 and 9 informative shape and statistical characters, respectively. Single trees of 251 steps (CI=0.785, RI=0.817; $g_1 = -0.916$; PTP test, $p = 0.001$), 138 steps (CI = 0.652, RI=0.628; $g_1 = -0.810$; PTP test, $p = 0.001$), and 394 steps (CI=0.728, RI=0.748; $g_1 = -0.909$; PTP test, $p = 0.001$) were recovered after exhaustive searches using shape characters (a), statistical characters (b), or both (c), respectively. Bootstrap values $>50\%$ are shown for individual nodes. The best RCC support tree is also shown together with total decay and cladistic information content (CIC). Numbers on branches of the RCC support tree indicate decay indices. The CIC for the entire tree was 13.34 bits. E, Ehrhartoideae; P, PAC-CAD clade; Po, Pooideae.

recovered using shape and statistical characters and branches were appropriately supported by bootstrap and DD analysis (Fig. 3). Phylogenetic trees showed that mRNA domains from rice (*O. sativa* L. and *O. brachyantha* L.) diverged earlier than domains from maize and two species in the Pooideae, rye grass (*Lolium perenne* L.) and barley (*Hordeum vulgare* L.).

Phylogenetic Analysis of rRNA Structure

Phylogenetic trees were also reconstructed from topological characters describing the secondary structure of one of the molecular components of the

Table 1. Alignment of shape characters describing *enod40* mRNA substructures

Character number	Characters	
	Shape	Statistical
Character number	11111111	11122222
	1234567890123456	789012345
Domain	1111112222333333	111222333
Type	sshhbbsshhbbshbb	QSPQSPQSP
<i>Glycine max</i>	I06032L264DA7300	A094AA2A8
<i>Medicago truncatula</i>	J0B033L0501M7500	5153A21A9
<i>Oryza sativa</i>	G07057G050359421	740000A69
<i>Oryza brachyantha</i>	H05057F050349411	06410347A
<i>Zea Mays</i>	H24451000009410	262A06000
<i>Hordeum vulgare</i>	P0504C000009411	2AA503000
<i>Lolium perenne</i>	N05065000009411	478103000

Note. Shape characters include number of nucleotides in coaxial stem tracts (s), hairpin loops (h), and unpaired segments in stem tracts (b). Statistical characters include the Shannon entropy (Q), mean size of stems (S), and base-pairing propensity (P). To accommodate the length of the stem tract, character 11 was weighted double.

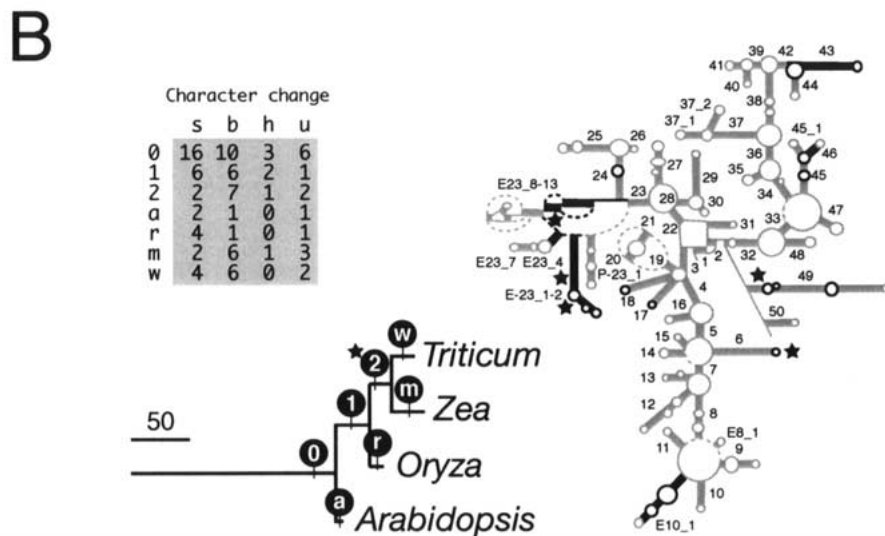
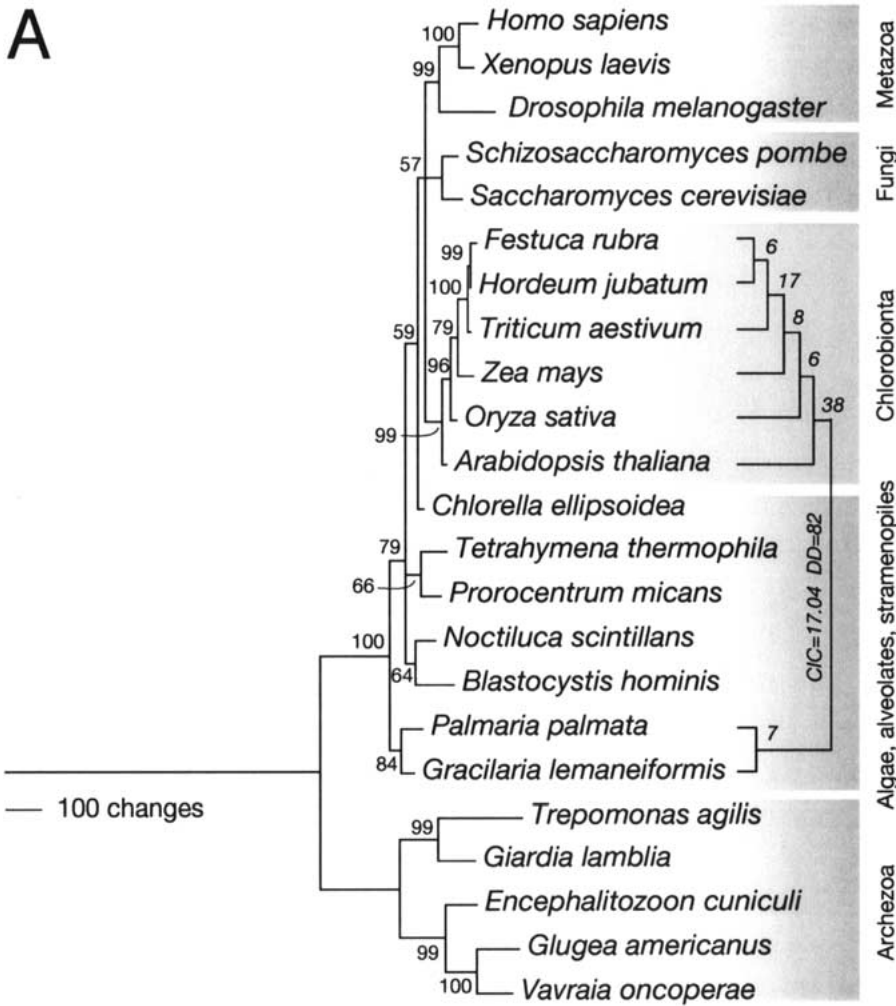
Table 2. Alignment of phylogenetically informative shape characters describing plant SSU rRNA structures

Type	Character number	
	Shape	Statistical
	1111111122222222222222222222222244445555555555555555555	122224466667455899999000012222245556669016811222335566777888
	990476713790956801458015680236895892370347817026031613124124	
Type	bbbbhbbsbsbbshsbbhsbbbussbsbsussusubshsuussbbsbbbbbbsbu	
<i>Festuca rubra</i>	040B156D1D711J1I114I11145A1A1544B149B71I6I87771766223113210F	
<i>Hordeum jubatum</i>	140B156D1D711J1I114I11145A1A1544B149B71I6I87771766223113210E	
<i>Triticum aestivum</i>	140B156D1D712H2I114I11145A1A1544B149B71I6I87771766112313211F	
<i>Oryza sativa</i>	104A005D1D712H2D064D20086B0B0635A059A81I6I87770766223101321F	
<i>Zea mays</i>	140B105D1D712H2E11BE50135B0B0534A149A80I6I78660777224509421F	
<i>Arabidopsis thaliana</i>	104A005E0E402H2K003K00055B0B0545A058A81J4J86662666223101321F	

Note. Coded characters represent number of nucleotides in stem tracts (s), hairpin loops (h), bulges (b), and unpaired segments (u).

ribosome, SSU rRNA (Table 2). Statistical characters were not analyzed because the structure of SSU rRNA is complicated and at present we can only reconstruct trees from ensembles of individual structural features in relatively simple molecules. A total of 23 molecules obtained from eukaryotic organisms spanning unicellular and multicellular organisms were analyzed, including molecules from six plant species. A wide range of outgroup molecular species was included to test the possible effect of long branches on phylogenetic reconstruction. There was significant phylogenetic signal in the data and the inferred trees had BS values of >90% for about half of the internal nodes (Fig. 4A). The tree showed the unprecedented diversity of the eukaryotic domain (especially evident in protoctist lineages) and groupings matched for the most part those inferred by comparative sequence analysis and traditional classification. The branching patterns of grass species were appropriately supported by bootstrap and DD analysis and were similar to those obtained with SRP RNA and *enod40* mRNA, with rice diverging earlier than maize and wheat. A RCC support tree derived

Fig. 4. Reconstruction of phylogenetic trees from the structure of SSU rRNA molecules. **A** Phylogenetic reconstruction of a eukaryotic tree from a total of 340 informative of 662 total RNA "shape" characters. Two trees of 3356 steps (CI=0.559, RI=0.647; $g_1 = -1.219$; PTP test, $p = 0.001$) were retained after a heuristic search with TBR branch swapping and 10 replicates of random addition sequence. The shape tree shown is congruent with the 50% majority-rule consensus. Bootstrap values (%) are shown for individual nodes with bootstraps over 50%. The most supported and plant-inclusive RCC support tree is also shown together with total decay and cladistic information content (CIC). Numbers on branches of the RCC support tree indicate decay indices. The CIC for the entire tree was 88.87 bits. **B** Phylogenetic tracing of character change in SSU rRNA structure. Character evolution was traced in a tree of selected plant species and variable structural characters were identified in the ribosomal molecule. SSU rRNA contains 50 universal helical stem segments, defined as those separated by multibranching or pseudoknot loops, and several segments specific to defined taxa. A schematic representation of the secondary structure of the ancestral eukaryotic molecule shows sequences drawn clockwise from the 5' to the 3' terminus, stem segments (bars sized as base pairs) corresponding to the 5' (helices 1–21), central (helices 22–31), and 3' domains (helices 32–50), and variable characters (highlighted in black). Changes in the branch corresponding to the ancestor of maize are identified with a star. A single tree of 294 steps (CI=0.888, RI=0.522; $g_1 = -0.867$; PTP test, $p = 0.001$) was retained following an exhaustive search. Changes are indicated in the table in B and were labeled with letters on the branches of the tree, together with bootstrap support values.



from DD analysis illustrates the robustness of these relationships (Fig. 4A). Character change was traced in the branches leading to these lineages and variable structural characters (Table 2) were identified in the SSU rRNA molecule (Fig. 4B). With one exception (stem 49), none of the changes occurred in structural

features that were important for the functioning of the ribosome, such as intersubunit bridges and tRNA binding sites. Instead, most were confined to segments known to be highly variable in the SSU rRNA molecules (Wuyts et al. 2001). Stem 49 is the penultimate helical segment of SSU rRNA that acts as a

Table 3. Alignment of phylogenetically informative shape and statistical characters describing grass ITS1 rRNA structures

Character number	Character	
	Shape	Statistical
	11111111122222222333333334555556666666	11
	123456789012345678123456789012345676012390123567	12345678901
Type	esBbhsBbmsBbhsBbmsHmsBbhsBbshmsBbhsBbshmsBbhsBbshms	<i>SLNNNDGGPQ</i> <i>sb1c eee</i>
<i>Amphipogon caricinus</i>	O000040040000H6LD000C233K344000000000000G3B40000	34A87285557
<i>Anisopogon arenaceus</i>	C0000514A3006B36E347C223E12600000000B3A400000000	357551A2347
<i>Aristida purpurea</i>	OF815720020000600J0008126C144000000000000O2440368	97254325656
<i>Brachyelytrum erectum</i>	9400I500H0000J3E70003004D22700000000E35400000000	37537271218
<i>Bromus inemis</i>	G000031490000600N34JD113H11400060030E11400000000	A9246191124
<i>Chusquea latifolia</i>	D0000500P5114300H000A143C124000E3C3G0000411140000	47878285638
<i>Eriachne trisetia</i>	00000300H5114300O000D125A274000600400000L5C40000	36747283429
<i>Leersia hexandra</i>	100005006000C24D549A004L48400000000Q6K400000000	52AA31A9AA3
<i>Lithachne humilis</i>	80000411N5114S6D70004005H29335E00000000000000000	267361A2315
<i>Micraira subulifolia</i>	D0000A25C5114G3390006004E384000000000000K3343000	449823678A4
<i>Microlaena stipoides</i>	6619760040000300L000D276G25400081430000000000000	77112251132
<i>Oryza sativa</i>	8000050093006C25F0007144R39337E00000000000000000	5732419232A
<i>Panicum bisulcatum</i>	F0000714E0000500I00040056004000B35504003N3440000	77264263435
<i>Pennisetum setaceum</i>	F0000714E0000500I00040056004000B35504003N3440000	77344271355
<i>Sporolobus airoides</i>	I000061473008D13F000C126D11400000000000J4A40000	77344271355
<i>Stipa ichu</i>	C0000514I0000700M000B116F35500060030E11400000000	794461A3327
<i>Thysalonaena maxima</i>	900002005500J500O000E233B16400000000000R3640247	AA356336747
<i>Zea mays</i>	C0000600B5004400I0007004F27400081430G23400000000	684461A4554

Note. Shape characters include number of nucleotides in coaxial stem tracts (s), hairpin loops (h), bulges in stem tracts (b), and unpaired segments in multi-loops (m) and external segments (e), and number of bulges (B). Statistical characters include (listed in order) the mean size of stems (*S*), and loops (*L*), mean number of stems (N_s), mean number of paired (N_b) and unpaired (N_u) nucleotides, number of components (N_c), mean loop degree (*D*), minimum free energy (*G*), minimum free energy of the thermodynamic ensemble (G_c), base-pairing propensity of the ensemble (P_c), and Shannon entropy (*Q*).

functional relay during protein synthesis (Cate et al. 1999). Changes in this central structural component can impact ribosomal dynamics and were here confined to the clade that encompasses maize and wheat (branches 2, w, and m; Fig. 4B). The exclusion of outgroup taxa did not affect the topologies of grass taxa (Fig. 4; data not shown). Cumulative exclusion of Archezoa, the Algae, Alveolates, and Stramenopiles group, Fungi, and Metazoa, in that order, did not affect the topology of angiosperm phylogeny, and exclusion of *A. thaliana* did not affect the topologies of grass taxa.

Phylogenetic Analysis of Spacer rRNA Structures

Phylogenetic trees describing the evolution of the ITS1 rRNA spacer were reconstructed from shape and statistical characters (Table 3). A group of 18 ITS1 sequences representative of major groups in Poaceae were included in the analysis. The most-parsimonious tree obtained using shape characters showed a late and separate branching of members of the Pooideae subfamily and PACCAD clade. Representative molecules from the Ehrhartoideae and Bambusoideae subfamilies were generally basal to the PACCAD lineages. However, molecules from *Chusquea latifolia* L.G. Clark and *Microlaena stipoides*

(Labill.) R.Br. diverged immediately after maize. Note that only few clades were well supported by bootstrap replication. However, DD analysis produced an all-inclusive basic RCC support tree with relatively high decay indices. Statistical characters carried little phylogenetic information and the resulting tree grouped the *Panicum bisulcatum* Thunb. and *Pennisetum setaceum* (Forssk.) Chiov. molecules together (56% BS) and separated reliably the early diverging *Leersia hexandra* Sw. ITS1 molecule from the rest (64% BS) (data not shown). Combination of shape and statistical characters did not alter topologies or improved their support. The phylogenetic relationships of representative molecules, including those from rice and maize, inferred using shape and statistical character data again suggested the early branching of the Ehrhartoideae (Fig. 5).

Tree Reconstruction and the Polarization of Structural Characters in RNA

Phylogenetic trees were always more parsimoniously reconstructed when the polarization of shape or statistical characters invoked the tendency toward order in RNA. For example, the *enod40* RNA trees reconstructed from characters polarized to-

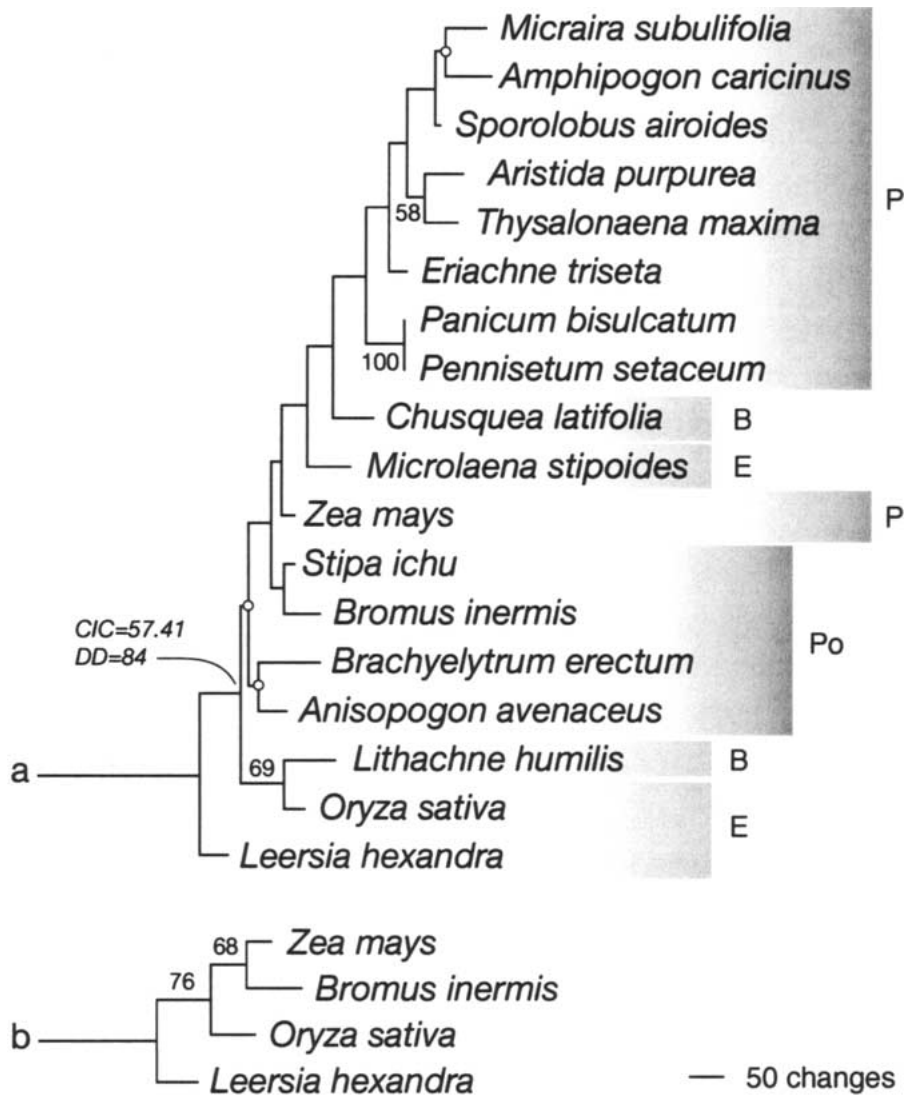


Fig. 5. Shape phylogeny reconstructed from the secondary structure of ITS1 spacer rRNA molecules. Phylogenies were inferred using unconstrained maximum parsimony as the optimality criterion in PAUP. Characters were equally weighted except for stems, which were weighted double to account for nucleotide number. A single most-parsimonious tree (1624 steps; CI=0.416, RI=0.534; $g_1 = -0.281$; PTP test, $p = 0.001$) was recovered from shape characters after a heuristic search with TBR branch swapping and 50 replicates of random addition sequence (a). Double decay analysis described a single most supported topology. Total

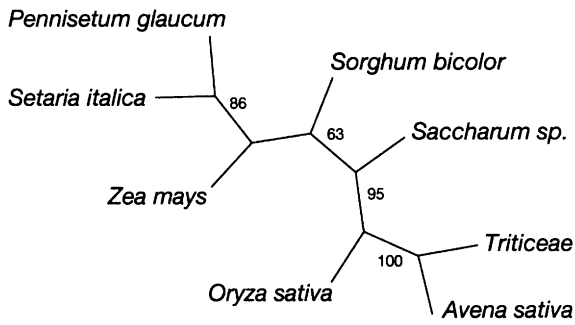
decay and CIC values are given for the RCC support tree matching the indicated clade. Nodes labeled with open circles have decay indices of four steps. All others have indices of six steps. A single most-parsimonious tree (569 steps; CI=0.763, RI=0.386; $g_1 = -0.281$; PTP test, $p = 0.001$) was recovered from combined shape and statistical character datasets after an exhaustive search (b). Bootstrap values (%) are shown for individual nodes with bootstraps over 50%. E, Ehrhartoideae; B, Bambusoideae; P, PACCAD clade; Po, Pooideae.

ward structural order were more parsimonious than trees reconstructed using characters that were polarized in the opposite direction. In this case, single trees were generated from shape, statistical, and combined character sets had 312 (CI=0.772, RI=0.796; $g_1 = -0.387$, PTP test, $p = 0.001$), 145 (CI=0.621, RI=0.604; $g_1 = -0.444$, PTP test, $p = 0.001$), and 461 (CI=0.718, RI=0.733; $g_1 = -0.416$, PTP test, $p = 0.001$) steps, respectively. In all cases these reconstructions were considerably less parsimonious than those described in Fig. 3.

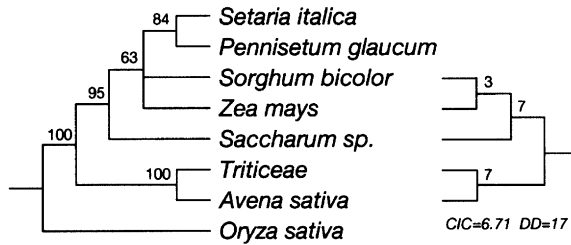
Phylogenetic Analysis of Large-Scale Chromosomal Rearrangements

The rooted phylogenies describing the evolution of RNA structure appear to indicate that the ancestors of rice diverged earlier than those of wheat and maize. In search of other characteristics that may reveal this evolutionary pattern, I generated phylogenetic trees from characters that describe major genomic rearrangements in “circular” maps that align grass genomes and characterize them in terms of “rice linkage blocks” (Fig. 6). These characters de-

A



B



C

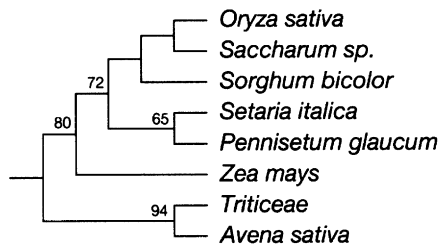


Fig. 6. Reconstruction of phylogenetic trees from chromosomal rearrangements. Phylogenies were inferred using unconstrained maximum parsimony as the optimality criterion from 95 informative characters describing translocations, inversions, and chromosomal distributions of linkage blocks. One most-parsimonious unrooted tree of 132 steps (CI=0.735, RI=0.720; $g_1 = -1.955$; PTP test, $p = 0.001$) was recovered after an exhaustive search using undirected characters (A). When characters were polarized by considering rice blocks “freestanding,” one most parsimonious tree of 132 steps (CI=0.735, RI=0.759; $g_1 = -1.521$; PTP test, $p = 0.001$) was recovered (B). In contrast, when characters were polarized in the opposite direction, a single tree of 195 steps (CI=0.564, RI=0.603; $g_1 = -0.876$; PTP test, $p = 0.001$) was obtained (C). Bootstrap values (%) are shown for individual nodes with bootstraps over 50%. Total decay and CIC values are given for the RCC support tree matching the indicated clade in the most-parsimonious rooted topology. The most supported RCC support tree is also shown together with decay indices (on branches), total decay and cladistic information content (CIC). The CIC for the entire tree was 17.04 bits.

scribe large-scale chromosomal inversions and translocations and instances of chromosomal orthology to more than one rice chromosome (Table 4). The genomes of Triticeae and oats (*Avena sativa* L.) represented the Pooideae in these studies and those of sugar cane (*Saccharum* species), maize, sorghum (*Sorghum bicolor* [L.] Moench), pearl millet (*Pennisetum glaucum* [L.] R. Br.), and foxtail millet (*Setaria italica* L.) represented the Panicoideae. A single most-parsimonious unrooted tree of 132 steps was recon-

structed using undirected characters (Fig. 6A). The tree showed two well-supported monophyletic groups representing the Pooideae and the Panicoideae. Within this last group, the millets were clearly monophyletic. I then compared this tree reconstruction with trees generated using directed characters. Characters were first polarized by considering chromosomal rearrangements (relative to rice, and necessary to describe present-day chromosomes) as derived states. The resulting tree was found to be equally parsimonious to the undirected reconstruction and showed that the most parsimonious location for attaching the hypothetical ancestor was the internode by which rice was connected to the tree (Fig. 6B). The relationships between ingroup taxa were not changed. Attachment of the all-cero hypothetical ancestor (displaying character homogeneity) to the ingroup showed absence of incongruence between the inferred ancestor and the possible ingroup nodes. In contrast, when characters were polarized in the opposite direction, the reconstructed tree was less parsimonious (195 steps) and its topology was changed (Fig. 6C).

Discussion

Congruent Evolution of Geometrical and Statistical Features in RNA Structure

In this study, I first focus on the evolution of RNA secondary structure, its study with cladistic methods, and the construction of intrinsically rooted phylogenies. RNA secondary structure can be considered a scaffold for three-dimensional architecture at atomic resolution. This was recently confirmed when rRNA structures predicted using the comparative sequence analysis method (James et al. 1989; Wuys et al. 2001) matched those derived from crystallographic studies (Yusupov et al. 2001). RNA structure is also a suitable fitness indicator in evolutionary studies *in silico* (Schuster et al. 1994, 1997; Fontana and Schuster 1998; Ancel and Fontana 2000). Fitness landscapes reflect the folding of RNA molecules and a genotype–phenotype map that has a statistically rich structure (Fontana 2002). Within the confines of these landscapes, molecules evolve in steps toward structural optimality by increasing conformational order and diminishing frustration by destabilizing loops. Since natural selection acts on molecules, we would expect direct evidence of links between thermodynamics and sequence evolution (Rivas and Eddy 2000; Washietl and Hofacker 2004). One approach has been to compare the structure of evolved and random sequences (Stegger et al. 1984; Higgs 1993, 1995; Schultes et al. 1999; Steffens and Digby 1999; Gulyaev et al. 2002). In these studies, evolved molecules

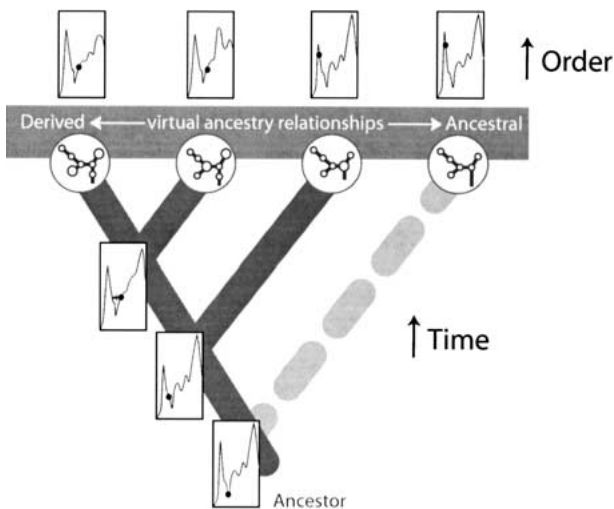


Fig. 7. A model for evolution of RNA structure. The cartoon illustrates in two dimensions how molecules traverse sequence space within a rugged multidimensional fitness landscape and are subject to evolutionary lock-in and diffusive mechanisms during the branching of lineages (cladogenesis). RNA molecules tend to reach optimal peaks of increased order and decreased frustration in a world of suboptimal evolutionary outcomes. Landscapes are illustrated with individual plots and the status of the adaptive walks indicated by a dot. The evolutionary interplay of mutation, natural selection, self-organization, and environment alters the landscapes and sometimes favors diffusive walks in neutral space (jumps between peaks illustrated with arrows connecting peaks). This permits a more efficient exploration of structural space but has the associated cost of diminished optimization relative to other molecules. Consequently, molecules that have a more ancestral origin (i.e., have not been the subject of many diffusive explorations) exhibit increased optimality and display structures that are more ordered. Phylogenies presuppose ancestors and inferences about ancestral molecular species are drawn from extant molecules and the model of character state transformation proposed. In practice, phylogenies are rooted by defining virtual relationships between extant molecules when connecting a hypothetical ancestral molecule to the base of the tree (dashed line depicting a basal branch). Landscapes describing ancestral molecules are provided at the internal nodes of the tree.

these by diffusive walks in neutral space. This has been recently demonstrated experimentally in elegant *in vitro* evolution experiments (Schultes and Bartel 2000). The existence of local optima and a rugged landscape defines a world of sub-optimal evolutionary outcomes in which conformation is not perfectly optimized and frustration is diminished by natural selection. These suboptimal conformations are here used in comparative exercise to infer phylogenetic relationships. Figure 7 also illustrates the rationale used to define ancestor–descendant relationships and root phylogenetic trees. Evolved structures are embedded in extant molecules and these are placed at the leaves of the tree. These structures derive from a common ancestor and consequently have been evolving at the same time. However, history is differentially imprinted on them as they transverse evolutionary landscapes and produce lineages. Consequently, phylogenies presuppose ancestors and ancestral nodes and inferences about them must be

derived from extant data and a model of character state transformation. Character polarization defines virtual ancestral–descendant relationships between the structure of molecules that are being studied, and these relationships can be established by connecting a hypothetical ancestral molecule that is also virtual to the base of the tree (Fig. 7). According to the model, molecules that have a more ancestral origin will exhibit increased optimality and display structures that are more ordered and less frustrated. In contrast, those that had the chance to explore structural space more efficiently (e.g., by diffusive walks in neutral space) should display suboptimal but more refined structures.

Grass Evolution Inferred from RNA Structure

In previous studies, I have shown that tree topologies were congruent with those derived from sequence and traditional classification at different evolutionary scales (Caetano-Anollés 2001, 2002a, b). The eukaryotic tree reconstructed from SSU rRNA (Fig. 4), for example, depicts (with exceptions) established classification, supporting (or failing to falsify) the assumption of character polarization. In this study, phylogenies reconstructed from grass SRP RNA, *enod40* mRNA, SSU rRNA, and ITS1 rRNA structure (Figs. 2–5) were generally in good agreement with each other and with those generated from their sequence (GPWG 2001; Hsiao et al. 1999; Hamby and Zimmer 1998). Furthermore, reconstructed trees were always more parsimonious than those generated by invoking a tendency toward disorder (see Results). Based on these considerations and the arguments put forth in the previous section, I determined a sequence of cladogenic events for the grass family without having to depend on outgroups.

The phylogeny of the grass family assembled by the GPWG (2001) combines 8 character sets, some of which support strongly (chloroplast restriction-morphological data), moderately (*rbcL*), and weakly (chloroplast restriction data and ITS rRNA) the existence of a Pooideae and PACCAD clade. Other character sets support strongly (molecular data pooled together), moderately (*rpoC2*), and weakly (*ndhF* and *phyB*) the existence of a BEP clade. Yet other sets are inconclusive, such as *waxy* and morphological characters. In particular, there were no morphological synapomorphies supporting the BEP clade. While the GPWG phylogeny of combined character sets suggests the BEP clade, the possibility that Pooideae is indeed a sister of the PACCAD clade cannot be ruled out. Constraining the data to accommodate this topology produced a tree that was only slightly less parsimonious (Wilcoxon signed rank test, $p < 0.406$; GPWG 2001). From these re-

sults, it appears that deep phylogenetic relationships are difficult to recover in the grasses. What was the sequence of cladogenic events during the early radiation that led to the Bambusoideae, Ehrhartoideae, and Pooideae lineages?

Phylogenies generated from the structure of all four classes of RNA molecules analyzed here support an early branching of representatives of the Ehrhartoideae subfamily (e.g., rice) and a late branching of representatives of the Pooideae (e.g., wheat) and of the PACCAD clade (e.g., maize) (Figs. 1–4). The GPWG phylogeny defines the Ehrhartoideae, Pooideae, and PACCAD clade as being monophyletic. In this study, these subfamilial groups are considered represented by the few species selected for analysis. Note, however, that only 22 grass genera are considered, of the over 700 described. Levels of support varied with the molecules analyzed, but were maximal with SSU rRNA and minimal with ITS1 rRNA. As expected, relationships were better supported by DD-derived RCC support trees than BS analysis. DD analysis determines decay indices of all n -taxon statements/partitions common to the most-parsimonious tree and produces a partition table that provides a more comprehensive summary of the strengths and weaknesses of phylogenetic hypotheses.

The inherently rooted phylogenies derived from SRP RNA, *enod40* mRNA, and SSU rRNA structure support the existence of a Pooideae and PACCAD sister clade relationship (Figs. 1–3). Structural models of these three molecular species have been inferred from X-ray crystallography, NMR spectroscopy, comparative analysis, and enzymatic and chemical probing. Phylogenies from spacer ITS1 rRNA placed some Ehrhartoideae and Bambusoideae species (including rice) as basal lineages and others (*Microlaena stipoides* and *Chusquea latifolia*) as members of a paraphyletic PACCAD clade (Fig. 5). The late divergence of these bambusoid and ehrhartoid grasses is an anomalous outcome, especially when comparing trees derived from the sequence and structure of spacer RNA molecules. The following considerations in order of importance can explain this outcome in spacer RNA: (1) branching relationships were poorly supported by the structural data, (2) a model of plant ITS1 structure has not been derived appropriately by positional covariance studies, and consequently anomalies could result from structural inaccuracies (tolerated here as systematic error) arising from incorrect folding of the spacer molecules, and (3) some nuclear spacer rRNA sequences may be divergent paralogues that escaped concerted evolution, a phenomenon that is common in angiosperms (Buckler et al. 1997). Despite these caveats, the monophyly of the Pooideae and its early divergence with respect to the paraphyletic PACCAD clade remain unquestioned and support the results

obtained with the other RNA molecules for which we know their structure.

Grass Evolution Inferred from Chromosomal Arrangements

In order to confirm the deep phylogenetic patterns inferred from RNA structure, I used a completely different approach. Rooted trees were reconstructed using characters that describe major genomic rearrangements of linkage blocks derived from the comparative genetic mapping of the grasses (Gale and Devos 1998a,b; Devos and Gale 2000). Large-scale insertions, translocations, and instances of chromosomal orthology to more than one rice chromosome were used to define phylogenetic relationships (Fig. 6). Only large-scale chromosomal rearrangements were included in the analysis, as these are expected to depict better the relationships between the major grass subfamilies. However, some species can accumulate more rearrangements than others (e.g., pearl millet [Devos et al. 2000] or species within Triticeae [Zhang et al. 1998]), but this may be a phenomenon linked to modes of reproduction and population size rather than deep phylogenetic evolutionary events. The strategy uses maximum parsimony to study positional homologies of large chromosomal segments. In this regard, it is comparable to several binary-encoding strategies used in the analysis of the order of gene complements (reviewed by Savva et al. 2003) but differs in the way characters are defined. For example, the position of genes has been coded by pairs of characters that describe genes flanking each gene or by single characters that encode adjacent gene pairs. Our coding scheme is mostly binary and defines the presence and absence of either a rearrangement, a set of two linkage blocks in an individual chromosome, or blocks orthologous to other chromosomes in each rice chromosome. This coding reflects a decision to use rice as the reference genome. The decision is arbitrary, since any ingroup genome could have been used as reference for this purpose. Using the approach, one most-parsimonious unrooted tree was reconstructed that showed two well-supported monophyletic groups representing the Pooideae and the Panicoideae (Fig. 6A). The fact that rice was not placed within one of these groups supports the tree topologies generated from RNA structures but is mute about the location of the root in the tree.

To address this issue, I compared the unrooted tree to rooted versions generated by polarizing character transformations in one of the two possible directions (Fig. 6B and C). The most-parsimonious tree of the two has the same length as the unrooted tree but places the ancestor of rice as the ancestor of the ingroup. The topology of ingroup genomes remains unchanged in this tree. In contrast, polarizing

character transformations in the opposite direction result in a tree of considerably larger length that groups rice with the Panicoideae. Consequently, the test places the ancestor of rice as the ancestor of the ingroup, supporting the claim that extant grass chromosomes can be considered combinations of linkage blocks of an ancestor of the rice genome (Devos and Gale 2000). My data cannot falsify this proposal. Interestingly, the comparative mapping of the genome of wild rice (*Zizania palustris* L.) revealed a linkage group configuration very similar to rice (e.g., with an apparent lack of rearrangement in linkage group 7) (Kennard et al. 1999), supporting again the ancestral nature of this genome organization. Note, however, that whole chloroplast genome comparisons of rice, maize, and wheat revealed a close relationship between the Ehrhartoideae and the Pooideae, which is at odds with the conclusions of this study (Matsuoka et al. 2002).

Conclusions

The evolutionary relationships inferred from RNA structure and genomes clarify the relationship of major clades in the grasses. The study also shows three findings that are of importance and are quite unexpected.

First, there was considerable phylogenetic signal embedded in the structure of polypeptide-coding *enod40* mRNA molecules, suggesting that strong evolutionary constraints are being actively imposed on mRNA structure. Functional RNA molecules possess well-ordered conformations and these could play crucial roles in defining biological function (Le et al. 2002). There are many well-documented cases where RNA loop-loop interactions appear to facilitate molecular recognition and trigger dynamic pathways (Brunei et al. 2002). One example involving mRNA is the formation of dimers during *bicoid* mRNA localization in the anterior pole of the *Drosophila* egg, a process that involves multiple steps during oogenesis and early embryogenesis (Takaiwa et al. 1985). Clearly, mRNA structure can be an important factor delimiting function. In the present case, *enod40* mRNA encodes two small polypeptides involved in the *Rhizobium*-legume symbiosis (Sousa et al. 2001; Rohrig et al. 2002) and apparently, act as structured RNA signals capable of regulating translation (Sousa et al. 2001). Second, characters that describe the geometry of RNA molecules carry similar phylogenetic signal than those describing molecular stability and plasticity. RNA geometry appears intimately linked to the thermodynamics of the system in an ongoing search toward increased order and decreased frustration. Finally, it appears that combinatorial exchange of chromosomal linkage blocks in the grasses also carries a strong

phylogenetic signal, suggesting large-scale rearrangement of genome structure can record organismal history. Interestingly, chromosome and RNA structural phylogenies were congruent. Their rooting and topology only matched when extant chromosomes were considered combinations of linkage blocks of an ancestor of the rice genome. These observations support previous claims on the evolution of grass genomes (Gale and Devos 1998b). As a corollary, the universal tendency toward order in RNA structure and the search for the most parsimonious organization of genome architecture appear to be mutually supported drivers of molecular evolution.

Acknowledgments. I gratefully acknowledge G.E. Caetano-Anollés for support and help in character coding, V. Knudsen, USIT, University of Oslo, for computer programs, and D. Pollock and an anonymous reviewer for valuable suggestions. Research was supported by CSREES-USDA-ILLU Hatch Project 802-316 and NSF Project MCB-0343126. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the fundings agencies.

References

- Ancel LW, Fontana W (2000) Plasticity, evolvability, and modularity in RNA. *J Exp Zool* 288:242–283
- Bennet MD, Leitch IJ (2001) Plant DNA C-value database (release 1.0, September 2001). Kew: Royal Botanic Gardens,
- Bennetzen JL, Kellogg EA (1997) Do plants have a one-way ticket to genomic obesity? *Plant Cell* 9:1509–1514
- Brunel C, Marquet R, Romby P, Ehresmann C (2002) RNA loop-loop interactions as dynamic functional motifs. *Biochimie* 84:925–944
- Bryant HN (1991) The polarization of character transformations in phylogenetic systematics: role of axiomatic and auxiliary assumptions. *Syst Zool* 40:433–445
- Buckler IV ES, Ippolito A, Holtsford TP (1997) The evolution of ribosomal DNA: divergent paralogues and phylogenetic implications. *Genetics* 145:821–832
- Caetano-Anollés G (2001) Novel strategies to study the role of mutation and nucleic acid structure in evolution. *Plant Cell Tissue Org Cult* 67:115–132
- Caetano-Anollés G (2002a) Evolved RNA secondary structure and the rooting of the universal tree. *J Mol Evol* 54:333–345
- Caetano-Anollés G (2002b) Tracing the evolution of RNA structure in ribosomes. *Nucleic Acids Res* 30:2575–2587
- Cate JH, Yusupov MM, Yusupova GZ, Earnest TN, Noller HF (1999) X-ray crystal structure of 70S ribosome functional complexes. *Science* 285:2095–2104
- Clayton WD, Renvoize SA (1986) *Genera Graminum: Grasses of the world*. Kew Bulletin Additional Series XIII. Royal Botanical Gardens, Kew. Her Majesty's Stationery Office, London
- Devos KM, Gale MD (2000) Genome relationships: the grass model in current research. *Plant Cell* 12:637–646
- Devos KM, Pittaway TS, Reynolds A, Gale MD (2000) Comparative mapping reveals a complex relationship between the pearl millet genome and those of foxtail millet and rice. *Theor Appl Genet* 100:190–198
- Farris JS, Källersjö M, Kluge AG, Bull C (1995) Testing significance of incongruence. *Cladistics* 10:315–319

- Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39:783–791
- Fontana W (2002) Modelling 'evo-devo' with RNA. *BioEssays* 24:1164–1177
- Fontana W, Konings DA, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. *Biopolymers* 33:1389–1404
- Fontana W, Schuster P (1998) Continuity in evolution: on the nature of transitions. *Science* 280:1451–1455
- Gale MD, Devos KM (1998a) Plant comparative genetics after 10 years. *Science* 282:656–659
- Gale MD, Devos KM (1998b) Comparative genetics in the grasses. *Proc Natl Acad Sci USA* 95:1971–1974
- Girard G, Roussis A, Gultyaev AP, Preij CWA, Paink HP (2003) Structural motifs in the RNA encoded by the early nodulation gene *enod40* of soybean. *Nucleic Acids Res* 31:5003–5015
- Gladyshev GP, Ershov YA (1982) Principles of the thermodynamics of biological systems. *J Theor Biol* 94:301–343
- GPWG (Grass Phylogeny Working Group) (2001) Phylogeny and subfamilial classification of the grasses (Poaceae). *Ann Missouri Bot Gard* 88:373–457
- Gultyaev PA, van Batenburg FHD, Pleij CWA (2002) Selective pressures on RNA hairpins in vivo and in vitro. *J Mol Evol* 54:1–8
- Hamby RK, Zimmer EA (1988) Ribosomal RNA sequences for inferring phylogeny within the grass family (Poaceae). *Plant Syst Evol* 160:29–37
- Hermann T, Patel DJ (1999) Stitching together RNA tertiary architectures. *J Mol Biol* 294:829–849
- Higgs PG (1993) RNA secondary structure: a comparison of real and random sequences. *J Phys I France* 3:43–59
- Higgs PG (1995) Thermodynamic properties of transfer RNA: a computational study. *J Chem Soc Faraday Trans* 91:2531–2540
- Hofacker IL, Fekete M, Stadler PF (2002) Secondary structure prediction for aligned RNA sequences. *J Mol Biol* 319:1059–1066
- Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte Chem* 125:167–188
- Hsiao C, Jacobs SWL, Chatterton NJ, Asay KH (1999) A molecular phylogeny of the grass family (Poaceae) based on the sequences of nuclear ribosomal DNA (ITS). *Aust Syst Bot* 11:667–688
- Huynen M, Gutell R, Konings D (1997) Assessing the reliability of RNA folding using statistical mechanics. *J Mol Biol* 267:1104–1112
- Huynen MA, Stadler PF, Fontana W (1996) Smoothness with ruggedness: the role of neutrality in adaptation. *Proc Natl Acad Sci USA* 86:397–401
- Jacobs BF, Kingston JD, Jacobs LL (1999) The origin of grass-dominated ecosystems. *Ann Missouri Bot Gard* 86:590–643
- James BD, Olsen GJ, Pace NR (1989) Phylogenetic comparative analysis of RNA secondary structure. *Methods Enzymol* 180:227–239
- Kauffmann SA (1993) *The origins of order*. Oxford University Press, New York
- Keenan RJ, Freymann DM, Stroud RM, Walter P (2001) The signal recognition particle. *Annu Rev Biochem* 70:755–775
- Kellogg EA (1998) Relationships of cereal crops and other grasses. *Proc Natl Acad Sci USA* 95:2005–2010
- Kellogg EA (2001) Evolutionary history of the grasses. *Plant Physiol* 125:1198–1205
- Kennard W, Phillips R, Porter R, Grombacher A (1999) A comparative map of wild rice (*Zizania palustris* L. $2n=2x=30$). *Theor Appl Genet* 99:793–799
- Kierzek E, Biala E, Kierzek R (2001) Elements of thermodynamics in RNA evolution. *Acta Biochim Polonica* 48:485–493
- Konings DAM, Gutell RR (1995) A comparison of thermodynamic foldings with comparatively derived structures of 16S and 16S-like rRNA. *RNA* 1:559–574
- Le SY Zhang K, Maizel JV Jr (2002) RNA molecules with structure dependent functions are uniquely folded. *Nucleic Acids Res* 30:3574–3582
- Maddison WP, Donoghue MJ, Maddison DR (1984) Outgroup analysis and parsimony. *Syst Zool* 33:83–103
- Maddison WP, Maddison DR (1999) *MacClade: analysis of phylogeny and character evolution*, Version 3.08. Sinauer Associates, Sunderland, MA
- Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters provides robust prediction of RNA secondary structure. *J Mol Biol* 288:911–940
- Matsuoka Y, Yamazaki Y, Ogihara Y, Tsunewaki K (2002) Whole chloroplast genome comparison of rice, maize, and wheat: implications for chloroplast gene diversification and phylogeny of cereals. *Mol Biol Evol* 19:2084–2091
- McCaskill JS (1990) The equilibrium partition function and base pair binding probabilities for RNA secondary structures. *Biopolymers* 29:1105–1119
- Page RDM, Holmes EC (1998) *Molecular evolution: A phylogenetic approach*. Blackwell Science, Oxford
- Petrov DA (2001) Evolution of genome size: new approaches to an old problem. *Trends Genet* 17:23–28
- Pollock DD (2003) The Zuckerkandl Prize: structure and evolution. *J Mol Evol* 56:375–376
- Rivas E, Eddy SR (2000) Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 16:583–605
- Rohrig H, Schmidt J, Miklashevichs E, Schell J, John M (2002) Soybean *enod40* encodes two peptides that bind to sucrose synthase. *Proc Natl Acad Sci USA* 99:1915–1920
- Rosenblad MA, Gorodkin J, Knudsen B, Zwieb C, Samuelsson T (2003) SRPDB: Signal recognition particle database. *Nucleic Acids Res* 31:363–364
- Savva G, Dicks J, Roberts IN (2003) Current approaches to whole genome phylogenetic analysis. *Brief Bioinform* 4:63–74
- Schultes EA, Bartel DP (2000) One sequence, two ribozymes: implications for the emergence of new ribozyme folds. *Science* 289:448–452
- Schultes EA, Hraber PT, LaBean TH (1999) Estimating the contributions of selection and self-organization in RNA secondary structure. *J Mol Evol* 49:76–83
- Schuster P, Fontana W, Stadler PF, Hofacker IL (1994) From sequences to shapes and back: a case study in RNA secondary structure. *Proc R Soc London Ser B* 255:279–284
- Schuster P, Stadler PF, Renner A (1997) RNA structures and folding: from conventional to new issues in structure predictions. *Curr Opin Struct Biol* 7:229–235
- Sousa C, Johansson C, CharonManyani H, Sautter C, Kondoroski A, Crespi M (2001) Translation and structural requirements of the early nodulin gene *enod40*, a short-open reading frame-containing RNA, for elicitation of a cell-specific growth response in the alfalfa root cortex. *Mol Cell Biol* 21:354–366
- Steffens W, Digby D (1999) mRNA have greater negative folding free energies than shuffled or codon choice randomized sequences. *Nucleic Acids Res* 27:1578–1584
- Stegger G, Hofman H, Fortsch J, Gross HJ, Randles JW, Sanger HL, Riesner D (1984) Conformational transitions in viroids and virusoids: comparison of results from energy minimization algorithm and from experimental data. *J Biomol Struct Dynam* 2:543–571
- Swofford DL (1998) *Phylogenetic analysis using parsimony and other programs (PAUP*)*, version 4.0. Sinauer Associates, Sunderland, MA

- Tacker M, Stadler PF, Bornberg-Bauer EG, Hofacker IL, Schuster P (1996) Algorithm independent properties of RNA secondary structure predictions. *Eur Biophys J* 25:115–130
- Takaiwa F, Oono K, Iida Y, Sugiura M (1995) The complete nucleotide sequence of a rice 25S rRNA gene. *Gene* 37:255–259
- Thiele K (1993) The holy grail of the perfect character: the cladistic treatment of morphometric data. *Cladistics* 9:275–304
- Thorley JL, Page RDM (2000) RadCon: phylogenetic tree comparison and consensus. *Bioinformatics* 16:486–487
- Washietl S, Hofacker IL (2004) Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol* 342:19–30
- Watson L, Dallwitz MJ (1992) *The grass genera of the world*. CAB International, Wallingford, Oxon, UK
- Wheeler WC (1990) Combinatorial weights in phylogenetic analysis. A statistical parsimony procedure. *Cladistics* 6:269–275
- Wilkinson M, Thorley JL, Upchurch P (2000) A chain is no longer than its weakest link: double decay analysis of phylogenetic hypothesis. *Syst Biol* 49:754–776
- Wright S (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proc Sixth Intl Congr Genet* 1:356–366
- Wuyts J, Van de Peers Y, De Wachter R (2001) Distribution of substitution rates and location of insertion sites in the tertiary structure of ribosomal RNA. *Nucleic Acids Res* 29:5017–5028
- Yusupov MM, Yusupova GZ, Baucom A, Lieberman K, Earnest TN, Cate JHD, Noller HF (2001) Crystal structure of the ribosome at 5.5 Å resolution. *Science* 292:883–896
- Zhang H, Jia J, Gale MD, Devos KM (1998) Relationships between the chromosomes of *Aegilops umbellulata* and wheat. *Theor Appl Genet* 96:69–75
- Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. *Science*, 244:48–52