J Mol Evol (2002) 54:333–345
DOI: 10.1007/s00239-001-0048-3

JOURNAL OF MOLECULAR EVOLUTION

Evolved RNA Secondary Structure and the Rooting of the Universal Tree of Life

Gustavo Caetano-Anollés

Laboratory of Molecular Ecology and Evolution and Division of Molecular Biology, Department of Biology, University of Oslo, N-0316 Oslo, Norway

Received: 13 September 2000 / Accepted: 27 August 2001

Abstract. The origin and diversification of RNA secondary structure were traced using cladistic methods. Structural components were coded as polarized and ordered multi-state characters, following a model of character state transformation outlined by considerations in statistical mechanics. Several classes of functional RNA were analyzed, including ribosomal RNA (rRNA). Considerable phylogenetic signal was present in their secondary structure. The intrinsically rooted phylogenies reconstructed from evolved RNA structure depicted those derived from nucleic acid sequence at all taxonomical levels, and grouped organisms in concordance with traditional classification, especially in the archaeal and eukaryal domains. Natural selection appears therefore to operate early in the information flow that originates in sequence and ends in an adapted phenotype. When examining the hierarchical classification of the living world, phylogenetic analysis of secondary structure of the small and large rRNA subunits reconstructed a universal tree of life that branched in three monophyletic groups corresponding to Eucarya, Archaea, and Bacteria, and was rooted in the eukaryotic branch. Ribosomal characters involved in the translational cycle could be easily traced and showed that transfer RNA (tRNA) binding domains in the large rRNA subunit evolved concurrently with the rest of the rRNA molecule. Results suggest it is equally parsimonious to consider that ancestral unicellular eukaryotes or prokaryotes gave rise to all extant life forms and provide a rare insight into the early evolution of nucleic acid and protein biosynthesis.

Key words: Cladistic analysis — Molecular evolution — Ribosomal RNA — Secondary structure — Universal tree

Introduction

The universal tree of life represents a hierarchical phylogenetic classification of the living world based on comparative analysis of sequences encoding ribosomal RNA (rRNA) and several proteins (Doolittle 1999). The currently accepted universal tree divides organisms in three domains, Archaea, Bacteria, and Eucarya (Woese 1987; Woese et al. 1990), and is currently rooted in the prokaryotic bacterial domain based on the evolution of paralogous proteins originated from ancient gene duplication (Gogarten et al. 1989; Iwabe et al. 1989; Doolittle 1999). The evolutionary tracing of these sequences has had a fundamental impact in evolutionary biology, providing benefits to fields as varied as protein chemistry, developmental biology and genomics and adding a temporal and dynamic component to the structure-function paradigm (Bull and Wichman 1998). However, the predictive ability of comparative sequence analysis to infer ancient phylogenies has been questioned on grounds of unequal rates of sequence evolution, mutational saturation, and long branch attraction artifacts (Philippe and Forterre 1999; Brinkmann and Philippe 1999). Moreover, the existence of lateral gene transfer threatens a

Correspondence to: G. Caetano-Anollés; 1320 Beacon Hill Lane, Knoxville, TN 37919, USA; *email:* gustavoc@mac.com

"natural" universal classification and the basis of deep phylogenetic analysis. Archaeal and bacterial genomes contain genes from multiple sources and share, for example, numerous housekeeping biosynthetic and catabolic genes (Woese 1998; Doolittle 1999). Similarly, rRNA genes can be completely replaced with corresponding genes from distant organisms, at least in bacteria, bypassing normal evolutionary processes (Asai et al. 1999). The formation of functional hybrid ribosomes in vivo suggests that co-evolution of rRNA and the translational ribosomal machinery may not completely preclude the horizontal transfer of rRNA. The universal tree is nevertheless supported by a massive database of rRNA sequence information. The choice of rRNA rests on it being ancient, central to translation, essential for cell function, and a highly social and conserved structure capable of interaction with hundreds of co-evolved protein and RNA molecules (Woese 1987; Green and Noller 1997). Unfortunately, because of concerted evolution, there are no paralogous genes that can root the rRNA universal tree.

RNA is ubiquitous and probably ancestral in origin to life (Gilbert 1986; Woese 1987; Joyce 1991). RNA molecules represent not only replicatable sequence genotypes but also selectable structural phenotypes with defined enzymatic and regulatory activity (Schuster et al. 1994). These features are generally mapped by folding algorithms (Zuker 1989; Schuster et al. 1997) or by positional covariance (i.e., patterns of correlated sequence substitution) in comparative sequence analysis (James et al. 1989; Gutell et al. 1994). RNA secondary structure is considered a scaffold for three-dimensional structure at atomic resolution and a suitable fitness indicator in evolutionary studies. In search of a theory that maps genotype, phenotype, and fitness to each other (Lewontin 1974; Kauffman 1993), structural changes in RNA have been modeled and the role of selection and self-organization traced (Higgs 1993; Huynen and Hogeweg 1994; Huynen et al. 1996; Fontana and Schuster 1998a; Schultes et al. 1999). The distribution of fitness values over genotype space defines a fitness landscape and its structure determines the peak-climbing evolutionary process (Wright 1932; Kauffman 1993). These landscapes have been used as evolutionary paradigms (Schuster et al. 1994; Huynen and Hogeweg 1994) to show that RNA populations evolve in discontinuous transitions interrupted only by long periods of neutral evolution (Huynen et al. 1996; Fontana and Schuster 1998a). During these periods, RNA molecules accumulate mutations freely while maintaining a dominant secondary structure. Despite advances, the evolution of molecular form in nature still remains vastly unexplored because of the need of an appropriate morphospace with which to compare evolved molecules by cladistic, phenetic, or statistical approaches. Here I focus on the evolution of RNA secondary structure (herein also referred to as "shape"), its study with cladistic methods, and its use to explore the intractable problem of the rooting of the tree of life. Phylogenies reconstructed from the shape of extant RNA molecules paralleled those derived from nucleic acid sequence, producing inherently rooted trees. This feature provided a tool to examine the RNA-based classification of the living world from a novel perspective and a rare insight into the early evolution of protein biosynthesis.

Materials and Methods

Sequence and structure. Small subunit (SSU) and large subunit (LSU) rRNA sequences were obtained from the Antwerp database (http://rrna.uia.ac.be), 5S rRNA sequences from http://cammsg3.caos.kun.nl, signal recognition particle (SRP) RNA from http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html, ribonuclease P RNA from http://www.mbio.ncsu.edu/RNaseP/home.html, snRNA from http://pegasus.uthct.edu/uRNADB/uRNADB.hml, tRNA from ftp://ftp.ebi.ac.uk/pub/databases/trna, and other sequences from GenBank. The aligned sequences and secondary structure information of rRNA, tRNA, and ribonuclease P RNA (derived from comparative sequence analysis) were downloaded as DCSE alignments or in special format. In the absence of structure-based alignments, sequences were aligned using CLUSTALX v. 1.64β (Thompson et al. 1997), alignments confirmed manually, and resulting data sets analyzed with parsimony methods. The secondary structures of these sequences were predicted using a free-energy minimization method (Zuker 1989) with updated energy parameters using the program *mfold* v. 3.1 (Mathews et al. 1999) or the web server (http://bioinfo.math.rpi.edu/~zuker/rna/form1.cgi).

LSU rRNA domains involved in peptidyl transferase function and the translational cycle were those defined by chemical footprinting, in vivo and in vitro functional studies (Green and Noller 1997) and high-resolution structural analysis in *Escherichia coli* 23S rRNA (Cate et al. 1999) and *Thermus thermophilus* (Yusupov et al. 2001). E-site interactions included nucleotides G2112, G2116, A2169, and C2394. A-site interactions included G1041, G1068, G1071, C1941, C2254, A2239, A2451, G2553, U2555, A2602, and U2609. P-site interactions included A1916, A1918, A1926, G2252, G2253, A2439, A2451, A2505, U2506, U2584, U2585, and A2602. Interactions (mostly on domain G, also known as V) defined 54 structural characters (18, 12, 10 and 14 E, A, P and P-A site characters, respectively) of which 42 were informative (Table 1).

Structural analysis. RNA secondary structures inferred from comparative sequence analysis or obtained using algorithms that minimize free energy were decomposed into substructural components and their features (such as the length of sequence tracts) were characterized using an alphanumerical format suitable for cladistic analysis. Homologous components were treated as discrete entities and analyzed with maximum parsimony methods. Other alternatives are possible. In related studies, substructural components were characterized by their thermodynamic stability measured using their minimum Gibbs free energy increments (Caetano-Anollés 2000, 2001). These values were treated as continuous quantitative characters and directly analyzed by restricted maximum likelihood methods, or were gap-recoded as discrete characters for maximum parsimony analysis.

RNA structures were first depicted as coarse-grained representations lacking information about the size of stacks and loops but adequate for quick identification of homologous structural characters, and homeomorphically irreducible tree (HIT) graphs (Fontana et al. 1993) that map (in the 5′-to-3′ direction) base pairs into internal nodes and unpaired bases into leaves. The nodes and leaves of the HIT graphs were used to code characters and construct data matrixes for cladistic

analysis. Coded characters were based on the length and number of double-helical stem tracts (S), hairpin loops (H), bulge and interior loops (B), and unpaired sequences (U) [such as free-ends and connecting joints (Tacker et al. 1996)], and other structural features [loop degree (D) and number of B loops in a stem tract (N), G:U base pairs, and modified nucleosides in tRNA]. Several coding schemes are possible, and two are illustrated in the examples of Fig. 1. Note that individual substructural components were characterized by one or more characters, each depicting an individual structural feature, and that structural features accommodated pseudoknots and non-canonical base pairing.

The coding of rRNA was based on secondary structure models for the large and small subunits inferred from sequences deposited in the Antwerp database and defined by comparative sequence analysis (van de Peer et al. 2000; Wuyts et al. 2001). The SSU model contains 50 universal stem tracts (S) (some missing in Archezoa) and several double-helical segments specific for Eucarya. The LSU model contains 100 universal stem tracts and several other stems specific to certain taxa. Note that universal stem tracts in these models are defined as those segments separated by multibranched or pseudoknot loops and are identified by numbers ordered in the 5′-to-3′ direction. Specific stems are named after the preceding universal stem followed by an underscore and a number. To illustrate the complexity in the structure and coding of rRNA, representations of the secondary structure model of the large subunit from *Sulfolobus solfataricus*, *Saccharomyces cerevisiae*, and an inferred ancestral molecule are shown in Fig. 2. The model of LSU rRNA is based on comparative sequence analysis and has been recently confirmed with the resolved (2.4 Å) crystal atomic structure of the archaebacterium *Haloarcula marismortui* molecule (Ban et al. 2000). The LSU molecule is treelike and is organized around a central multibranched loop that is closed by a stem helix (segment A1) joining the 5′ and 3′ ends of the molecule in Bacteria and most Archaea but absent in Eucarya. A total of nine domains (A-I) branch from the central loop. These domains are composed of up to 28 universal stem tracts, each separated from each other by unpaired sequences defining multibranched loops (U), containing in some cases numerous bulges and interior loops (B), and generally (74%) ending in a hairpin loop (H). Coded characters in both the SSU and LSU rRNA molecules were based on the length in nucleotides of distinguishing structural features (S, B, H, and U), and could be easily identified in DCSE alignment outputs. Character states were limited by the maximum number accepted by the phylogenetic analysis program (32 states) and were represented by the numbers 0–9 and letters A-V. Structural features with longer nucleotide lengths were given the maximum state, and if missing, the minimum state (0). Structural alignments listed characters characterizing the structure in the 5′-to-3′ direction as it is read in the sequence, and for each sequence segment, in the order S, B, H, and U. Stem tracts were defined by two complementary sequence segments and characters (named by a number and its prime) to account for the difference in nucleotide number between stem and unpaired tracts. Table 1 shows an alignment matrix of LSU rRNA structural characters containing tRNA-binding sites for 35 taxa covering the three domains of life. This is only a subset of the total data matrix for the LSU rRNA molecule, which can be retrieved from the TreeBASE repository (http://herbaria.harvard.edu/treebase/) under study and matrix Accession Numbers S053 and M1025, respectively.

In this study, topographic correspondence was a main criterion for determining character homology. Molecular topography involved three steps: (a) mapping of structural features in the context of the whole molecule, (b) proper encoding of characters, and (c) determining that encoded features were true homologies acquired from a common ancestor. The process uses a method for character analysis (Neff 1986) in which homologous features are first delimited by similarities (e.g., S_2 is in between S_1 and S_3) and hierarchies (e.g., S_1–S_3 belong to a same domain, and B_2 and H_2 are intricately related to S_2), and are then subjected to hypothesis testing. In the presence of a model inferred by positional covariance in sequences from a representative group of organisms, decisions to code homologies were simpler than those in aligned sequence and could be done manually or using coarse alignment tools. In the absence of structure-based alignments obtained from comparative sequence analysis, homology hypotheses required testing by iterative rounds of phylogenetic reconstruction. In this case, coarse-grained representations of structure were very useful for the identification and tracing of characters.

Character coding disregards information and implications of high-order structure [such as the establishment of complementary surfaces (e.g., U-turns), specific ion associations, and pseudoknot constraints; Draper 1996], coarse-graining its three-dimensional complexities into a simple framework of non-interacting helical segments. Coding relies however on correct prediction of secondary structure. While there is not yet a comprehensive understanding of sequence-structure relationships, comparative sequence analysis has been successful in predicting structures that were congruent with those determined by NMR and crystallography and thermodynamic-based folding algorithms have become increasingly more effective in matching structures inferred by positional covariance (Draper 1996; Schuster et al. 1997). Structural inaccuracies are therefore assumed not to be severe and are here tolerated as systematic error, provided structures result from a same comparative sequence study or are folded using a same algorithm. Interestingly, phylogenetic reconstruction using small structures (e.g., Y1 RNA) folded with the *mfold* or *RNAfold* (Vienna RNA package, ftp://ftp.itc.univie.ac.at/pub/RNA) programs appeared independent of folding algorithm. This suggests that any bias introduced by mapping sequences into shapes distributes uniformly in phylogenetic trees.

Phylogenetic analysis. Phylogenetic relationships were inferred using PAUP* v. 4.0 (Swofford 1999) and character reconstruction implemented in MACCLADE v. 3.08 (Maddison and Maddison 1999). Characters had multiple discrete states, were linearly ordered, and were polarized by fixing the direction of character state change using a transformation sequence that distinguishes ancestral states as those more stable thermodynamically (e.g., larger S and lower H, B, U, N, and D state values). The model of character state transformation is based on the hypothesis that evolved RNA molecules are optimized to produce highly stable folded conformations. This optimization process increases favorable and decreases non-favorable inter- and intramolecular interactions, and restricts alternative outcomes of the folding process. The hypothesis results in the polarization of structural characters in one out of two possible directions, a proposal that is supported by establishing that: (a) phylogenetic trees reconstructed from the secondary structure of RNA molecules exhibited inherently rooted topologies matching those known from systematic studies at widely different taxonomical levels (this study), (b) molecular evolution enhances conformational order over that intrinsically acquired by self-organization (Higgs 1993; Schultes et al. 1999; this study), and (c) thermodynamic principles generalized to account for non-equilibrium conditions can be used to verify a molecular tendency towards order and stability (e.g., evolutionary increase of the volume density of the Gibbs free energy in plant and animal development; Gladyshev and Ershov 1982) proposed in a thermodynamic-based theory of evolution (Gladyshev 1978; Black 1978; Gladyshev and Ershov 1982).

The data was encoded in the NEXUS format. The ANCSTATES command was invoked to define the ancestral states and polarize character change. Hypothetical ancestral molecules (Anc) were chosen as those having maximum stem lengths and absence of unpaired destabilizing regions. However, tree topologies remained unaltered when minimum states in unpaired regions were defined as ancestral. Phylogenetic trees were generally reconstructed using maximum parsimony as the optimality criterion. Generated trees were automatically rooted at the point where the hypothetical ancestor connected to the tree. Phylogenetic reliability was evaluated by the nonparametric bootstrap (BS) method (generally implemented using 10^3 replicates) and by decay analysis. The structure of phylogenetic signal in the data was tested by the skewness (g_1) 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

Table 1. Alignment of structural characters encoding tRNA-binding sites in LSU rRNA[a]

Character number	3333333555555556666666666666666666666777777777777 7777777
	2333333777779992333333333334444444777712333 4444 4 5555555 5
	9012345456790129012345677901234560127733454567 7 0124567
Character type	SBBHSBBSHSSBHSBSBBBBBBBBHSBBBBBBBSHSSSUUSBSBHSBSBBSUSB
Binding site	AAAAAAAPPPAAAAAEEEEEEEEEEEEEEEEEEEEEMMMMMPPPPMMMMMMPPPMMMM
Helix	D_{18} E_{26} E_{27} G_4 G_6 G_8 G_{17} G_{19} G_{20}
	$D_{18'}$ $E_{26'}$ $E_{27'}$ $G_{4'}$ $G_{6'}$ $G_{8'}$ $G_{19'}$ $G_{17'}$ $G_{20'}$
Desulfurococcus mobilis	6209602676606600Q000B26004Q03C20004341179C04054 0 C22747 1
Sulfolobus solfataricus	6209602676606600Q000B26004Q03C20004341179C04054 0 C22747 1
Halobacterium halobium	6209602676606600Q100B29104Q23C20014341179C04054 0 C22747 1
Haloferax volcani	6209602676606600Q100B29104Q23C20014341179C04054 0 C22747 1
Methanobacterium therm.	6209602676606600Q100B26004Q03C20014341179C04054 0 C22747 1
Methanococcus jannaschii	6209602676606600Q000B26004Q03C20004341179C04054 0 C22747 1
Chlamydia sp.	7109701676516516 51Q119B04Q03C20015351179C05055 0 C22747 1
Anacystis nidulans	6119611676516 51Q100B26004Q03C20015351179C05055 0 C22747 1
Synechocystis sp.	7109701676516 51Q100B26004Q03C20015351179C05055 0 C22747 1
Mycobacterium tub.	7109701676516 51Q000B26004Q03C20005351179C05055 0 C22747 1
Bacillus subtilis	7109701676516 51Q100B26004Q03C20015351179C05055 0 C22747 1
Bradyrhizobium japonicus	7109701676516 51Q000B26004Q03C20005351179C05055 0 C22747 1
Pseudomonas cepacia	7109701676516 51Q010B26004Q03C21005351179C05055 0 C22747 1
Helicobacter pylori	7109701676516 51Q000B26004Q03C20005351179C05055 0 C22747 1
Escherichia coli	7109701676516 51Q010B26004Q03C21005351179C05055 0 C22747 1
Thermus thermophilus	7109701676516 51Q000B26004Q03C20005351179C05055 0 C22747 1
Borrelia burgdorferi	7109702676516 51Q000B26004Q03C20005351179C05055 0 C22747 1
Thermotoga maritima	7109701676516 51Q000B26004Q03C20005351179C05055 0 C22747 1
Aquifex aeolicus	7109701676516 51Q000B26004Q03C20005351179C05055 0 C22747 1
Drosophila melanogaster	5309503676506 50Q000A20004Q00C10104341 19AC04154 1 C22747 1
Xenopus laevis	5309503676506 50Q000A20005Q31C10104341 19AC04154 1 C22747 1
Homo sapiens	5309503676506 50Q010A201IGQ60C11104341 19AC04154 1 C22747 1
Caenorhabditis elegans	5309503676506 50P000A20004P00C10104341 19AC04154 1 C22747 1
Saccharomyces cerevisiae	5309503676506 50O000A20004O00C10104341 19AB14154 1 B50747 1
Schizosaccharomyces pom.	5309503676506 50Q000A20004Q00C10104341 19AC04154 1 C22747 1
Oryza sativa	5309503676506 50P000A20004P00C10104341 19AC04154 1 C22747 1
Arabidopsis thaliana	5309503676506 50O000A20004O00C10104341 19AC04154 1 C22747 1
Brassica napus	430B403676506 50O000A20004O00C10104341 19AC04154 1 C22747 1
Tetrahymean thermophila	5309503676506 50O000A20004O00C10104341 19AC04154 1 C22747 1
Dictyostelium discoideum	5309503676506 50O100A20004O00C10114341 19AC04154 1 C22747 1
Prorocentrum micans	5309503676506 50O000A20004O00C10104341 19AC04154 1 C22747 1
Physarum polycephalum	5309503676506 50P011A20004P00C11103331 19AC03153 1 C22747 1
Zea mays (mitochondria)	6009602676 40A40L110B2000EL00C20115351 179C05055 0 C22747 1
Oryza sativa (plastid)	6119611676 40A40Q110B26004Q30C20115351 179B15055 0 B32747 1
Zea mays (plastid)	6119611676 40A40Q100B26004Q20C20015351 179B15055 0 B32747 0
Ancestral states	7000700606600 60Q000000000Q000000050 51100C05005 0 C00707 0

[a]Characters represent stems (S), bulges and interior loops (B), hairpin loops (H), and unpaired sequences (U), and are a subset of the 848 characters that characterize the complete LSU molecule They encode P, A, shared P and A (listed as M), and E binding sites. Universal helix tracts are identified (see Fig. 2).

modified Michevich-Farris index of incongruence among data sets and 10^3 heuristic search replicates (Farris et al. 1995). Topological congruence was measured using several tree comparison metrics (e.g., partition distance, PD, symmetric difference, SD, and strict joint assertions, SJA, from quartet analysis) and randomization tools implemented in COMPONENT v. 2.0 (Page 1993), and using pairwise consensus fork indices (CFI) (Fox et al. 1999). CFI is a conservative metric unblased by tree topology that measures the proportion of subclades (internal nodes) shared between dadograms [ranges from 0 (no identity) to 1 (total identity)].

Morphospace analysis. The metrics of base-pairing propensity (P), mean length of helical stem (S), and the Shannon entropy of the base-pairing probability matrix (Q), normalized to sequence length, were used to define a structural morphospace for RNA and to analyze evolved and randomized molecules, as described by Schultes et al. (1999). Random sequence cohorts from evolved RNA molecules (20 replicates each) were generated by permutation using heteropolymer randomization algorithms (*stringgen* and *omrokgen* programs available from V. Knudsen, USIT, Univ. of Oslo).

Results and Discussion

Phylogenetic Reconstruction of Molecular Shape

The origin and diversification of RNA secondary structure were traced using cladistic methods. Molecular structures were decomposed into substructural compo-

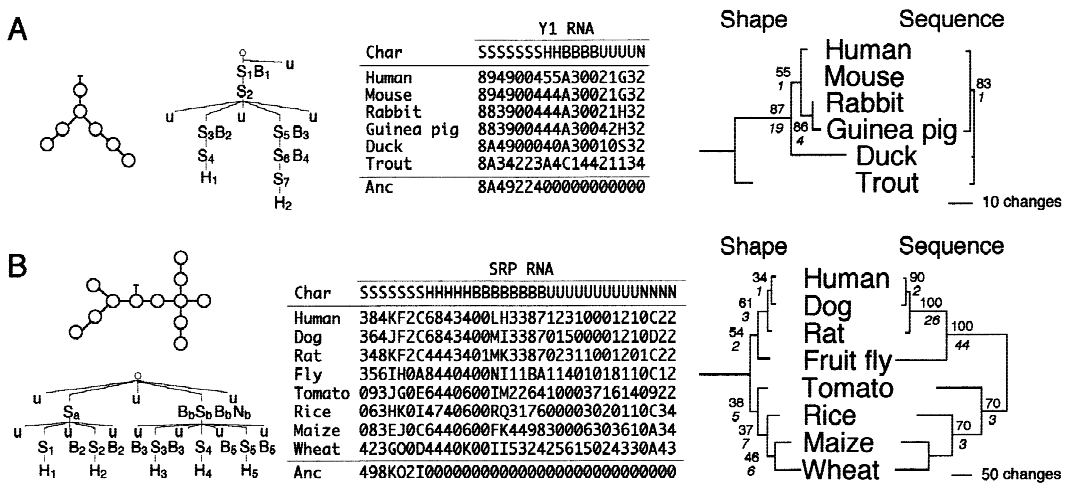

Fig. 1. Construction of shape phylogenies from small RNA molecules. The secondary structures of vertebrate Y1 RNA (107–112 nucleotide components of the Ro ribonucleoprotein complex) and bilateria-angiosperm SRP RNA (300–303 nucleotide components of cytosolic ribonucleoproteins) were inferred by folding and comparative sequence analysis, respectively. Coarse-grained structures were depicted as collections of loops (open circles) connected by double-stranded stems (lines). Structures were also represented as HIT graphs, with nodes and leaves defining the structural characters (char) used to construct a data matrix for cladistic analysis (see Materials and Methods) and polarized by including an hypothetical ancestor (Anc). For SRP RNA, characters were equally weighted except for stems that were weighted double to account for nucleotide number (S_b was weighted 6 because PAUP* accepts only 32 states per character). Phylogenies inferred from RNA structure or sequence data using unconstrained maximum parsimony in exhaustive searches were compared and found congruent (Y1 RNA, $p = 0.635$; SRP RNA, $p = 0.829$). Two rooted most-parsimonious Y1 RNA shape trees of 86 steps (CI = 0.860, RI = 0.739; $g_1 = -1.962$; PTP test, $p = 0.019$) and six sequence trees (seven steps, CI = 0.857, RI = 0.750; $g_1 = -0.713$; PTP permutation test, $p = 0.750$) were recovered. Trees shown were congruent with the 50% majority-rule consensus. Analysis of SRP RNA produced a shape tree of 488 steps (CI = 0.637, RI = 0.452; $g_1 = -0.425$; PTP test, $p = 0.018$) and a sequence tree of 325 steps (CI = 0.830, RI = 0.833; $g_1 = -0.744$; PTP test, $p = 0.001$). BS values and decay indexes (in *italics*) are shown for individual nodes.

nents such as double helical stems and unpaired regions. Homologous components were then coded as ordered and polar multi-state characters, following a model of character state transformation in which structures with increased thermodynamic stability were defined as being ancestral (plesiomorphic). Finally, the evolutionary history of molecular shape was reconstructed using maximum parsimony methods.

The strategy uses structure to infer phylogeny and therefore differs from comparative sequence analysis, where structure is inferred from phylogenetic diversification. This should not be construed as a circular argument because positional covariance focuses on conserved structural elements while "shape phylogenies" are based on structural features that are variable.

Several classes of functional RNA were analyzed at various taxonomical levels. RNA shape phylogenies were reconstructed from SRP RNA, Y RNA, ribonuclease P RNA, snRNA, tRNA, SSU rRNA, LSU rRNA, 5S rRNA, and spacer rRNA. With the exception of 5S rRNA, there was considerable phylogenetic signal in the data. Distribution of cladogram lengths and PTP tests showed the existence of strong cladistic structure ($p < 0.01$ and $p = 0.001$–0.019, respectively). Many clades were well supported by BS and decay analysis. Moderately supported shape phylogenies were even obtained from RNA molecules with short uninformative sequences. Figure 1 shows examples with structures defined by minimum free energy (MFE) folding (Y1 RNA) or comparative sequence analysis (SRP RNA). The shape trees matched traditional classification despite structures being coded by only 18 and 34 characters, respectively, and in the case of vertebrate Y1 RNA, presence of negligible phylogenetic signal in the short sequences examined. These examples show different coding schemes and illustrate the cladistic method employed.

Shape and sequence phylogenies were compared and found congruent at different evolutionary scales, showing it constitutes a general phenomenon. This is illustrated when analyzing minimum free-energy structures obtained from spacer rRNA molecules of fungal and plant origin. For example, the ITS1 spacers of soil-borne *Rhizoctonia solani* fungal isolates belonging to anastomosis group 4 were arranged in three groups according to habitat and virulence (Fig. 3A). This same pattern was observed in the analysis of the entire ITS sequence (Boysen et al. 1996). At the species level, a cross-section of fungi of the genus *Discula* that cause anthracnose in broadleaf temperate trees were grouped in four monophyletic clades corresponding to major species using both sequence and shape analysis of spacer rRNA (Caetano-Anollés 2001). Similarly, molecular shape relationships in the legume genus *Glycine* (Fig. 3B) matched sequence and genomic comparison (Kollipara et al. 1997). However, trees were rooted in wild perennial species while the cultivated soybean (*Glycine max*) and its wild annual progenitor (*Glycine soja*) were evolutionary

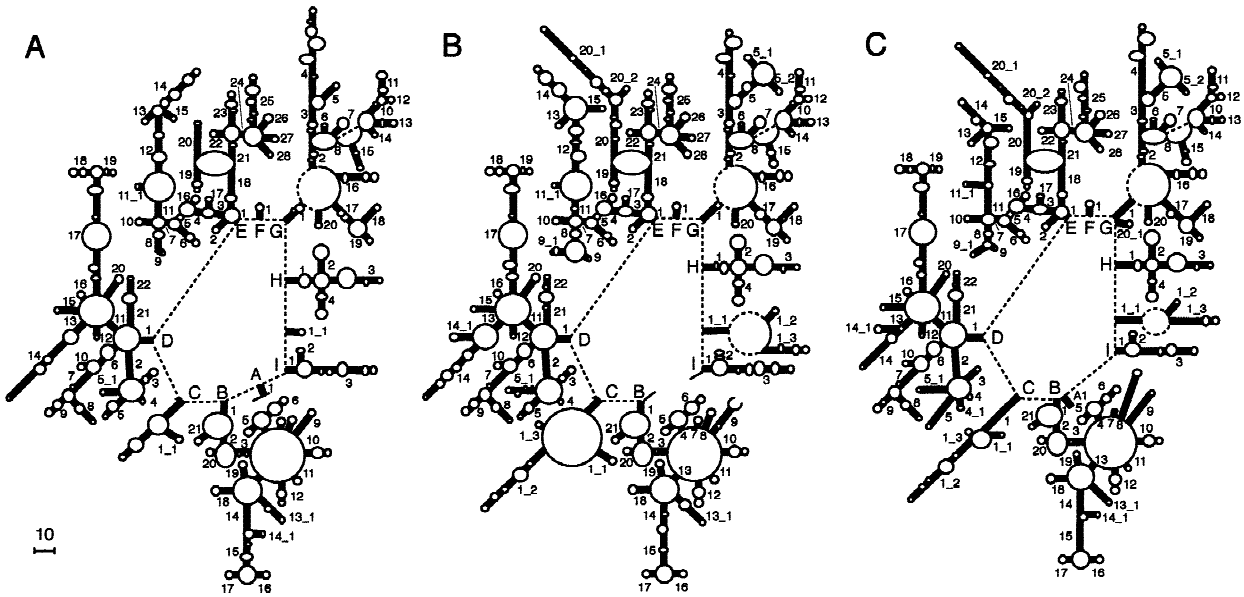

Fig. 2. Schematic representation of the secondary structure model of the LSU rRNA from representative prokaryotic and eukaryotic organisms and a putative universal ancestor. The sequences are drawn clockwise from the 5′ to the 3′ terminus, and helical segments are numbered in the same order and for each individual structural domain (A-I) (De Rijk et al. 1995; Wuyts et al. 2001). **(A)** *Sulfolobus solfataricus* subunit. **(B)** *Saccharomyces cerevisiae* subunit. **(C)** Ancestral molecule inferred by phylogenetic analysis of secondary structure. The ancestral molecule is a hypothetical contemporary entity that in the course of evolution was trapped in local optima of rugged adaptive landscapes (constrained by the mapping of sequence into structure). The scale measures stem tract length.

derived. Congruence was also evident at higher taxonomic level when analyzing grass species from the Pooideae, Oryzoideae, and Panicoideae subfamilies (Fig. 3C). Shape phylogenies obtained from ITS1 matched those inferred from the complete ITS sequence and agreed with accepted classification (Hsiao et al. 1994). In all cases, the null hypothesis of congruence could not be rejected when combined data sets were tested for homogeneity of data partitions ($p = 0.544$–0.999) and when fitting characters in topology-dependent PTP randomization tests ($p = 0.086$–1.000). Furthermore, tree topologies were generally in good agreement (PD = 0–11, SD = 0–0.429, SJA = 0–0.193, CFI = 0.67–1.0, for 6–14 leaved trees), rejecting the occurrence of taxonomic congruence by chance ($p < 0.01$). At small evolutionary scale, however, both sequence and shape phylogenies were globally poorly resolved, and congruence was sometimes a consequence of low resolving power (e.g., Y1 RNA, SD = 0.429, SJA = 0; *Glycine* ITS1, SD = 0.386, SJA = 0.063).

Homoplasy levels measured for example by the consistency index (CI) were usually higher in shape than in sequence data sets. This increase in cladistic noise can be explained by the fact that secondary structure is more prone to convergence and reversion than primary sequence. Mapping of RNA sequence into secondary structure produces dynamic and very rugged fitness landscapes, with small changes in genotype resulting in large changes in phenotype (Fontana et al. 1993; Huynen et al. 1993). Convergent evolution and homoplasy are enhanced in these adaptive landscapes as sequences and structures converge and diverge during adaptive walks in the rugged multipeaked terrain (Kauffman and Johnsen 1991; Kauffman 1993). Branching phylogenies climbing fix or deforming landscapes therefore reflect the landscape structure and high-dimensional space of molecular shape.

Character Polarization and Rooting of Trees

Shape trees are reconstructed using a “direct” method (*sensu* Nelson 1973) that polarizes entire topologies intrinsically, i.e. without recourse to outlying group comparisons or external pre-existing hypotheses of relationship. The model of character state transformation assumes a priori that evolution of individual substructures is independent from each other and results in molecules with increased conformational order and stability. The latter assumption was challenged using a comprehensive molecular morphospace (Schultes et al. 1999) that tested if evolved sequences were more “ordered” than randomized derivatives. Order was inferred from the measurement of three features that characterize stability (P and S) and uniqueness of folded conformations (Q) (Schultes et al. 1999). The values P and S are good indicators of how extensively folded and ramified are molecules analyzed. The entropy value Q is determined from the equilibrium partition function that measures the thermodynamic likelihood of base pairing during the energy minimization process of RNA folding. Evolved sequences had significantly larger P and S, and lower Q

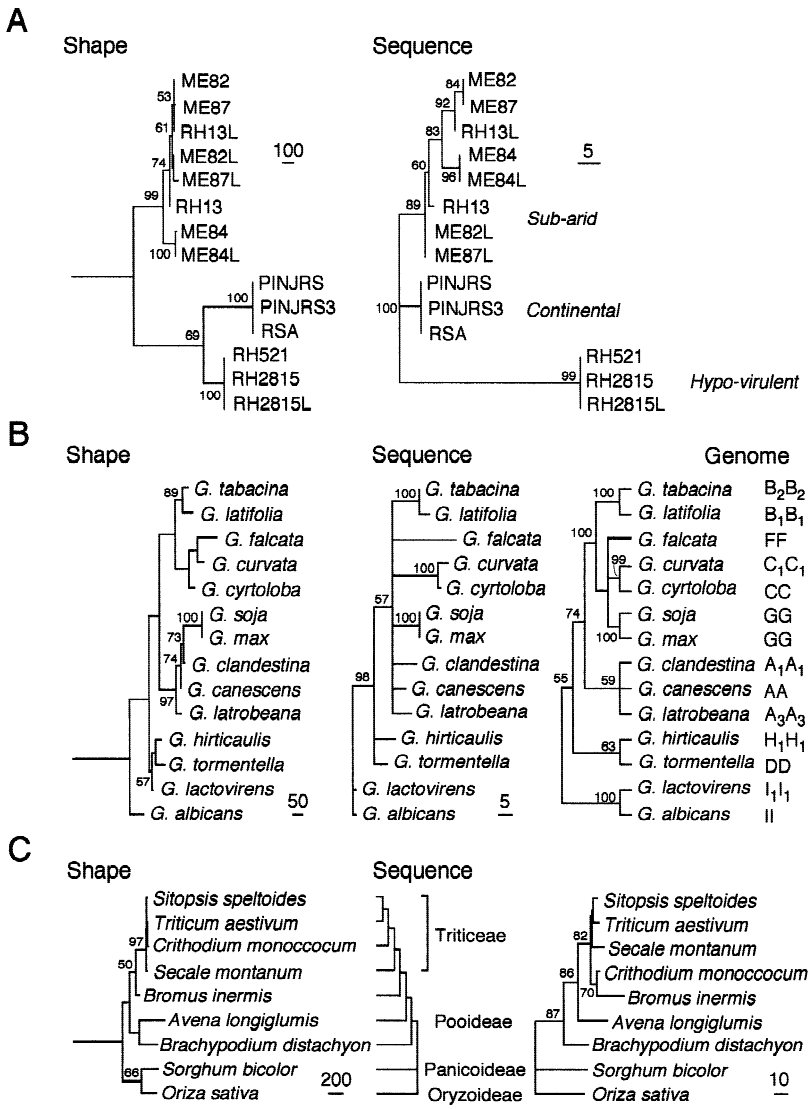

Fig. 3. Reconstructing historical lineage at different evolutionary levels from the shape and sequence of spacer ITS1 rRNA molecules. **(A)** Phylogenetic analysis of spacers from *Rhizoctonia solani* fungi recovered eight shape trees (2447 steps; CI = 0.745, Rl = 0.976; g_1 = −0.551; PTP test, p = 0.001) and one sequence tree (67 steps; CI = 0.955, RI = 0.979; g_1 = −2.011; PTP test, p = 0.001) in branch-and-bound searches. The shape tree shown is congruent with the 50% majority-rule consensus. **(B)** Analysis of spacers from *Glycine* legume species recovered a shape tree (1401 steps; CI = 0.542, RI = 0.664; g_1 = −0.627; PTP test, p = 0.001) and 25 sequence trees (108 steps; CI = 0.714, RI = 0.765; g_1 = −0.245; PTP test, p = 0.001) in branch-and-bound searches. The sequence tree shown is congruent with the 50% majority-rule consensus. Shape and sequence trees match a tree reconstructed from the entire ITS region that shows genomic relationships (symbols) established using cytogenetic, biochemical and molecular criteria (Kollipara et al. 1997). **(C)** Analysis of spacers from several grass species recovered single shape (2742 steps; CI = 0.604, RI = 0.494; g_1 = −0.492; PTP test, p = 0.015) and sequence (103 steps; CI = 0.718, RI = 0.580; g_1 = −0.692; PTP test, p = 0.001) trees in exhaustive and branch-and-bound searches, respectively. A free obtained from the overall ITS region is also shown. BS values >50% are given above nodes.

values (p = 0.05) than their corresponding permuted cohorts (and overall means) obtained by sequence randomization (data not shown). Results follow those recently reported for several classes of functional RNA (Schultes et al. 1999), confirming that the evolutionary process deforms the folding energy landscape towards higher order and stability. Therefore, considerations in statistical mechanics of evolved RNA molecules and correct experimental reconstruction of known rooted topologies (shown in section above) provide strong support to the validity of character polarization and the rooting of trees.

Structural rRNA Diversification in Archaea

Direct comparison of shape and sequence trees validated phylogenetic reconstruction of molecular shape at rather small evolutionary scale. To examine the usefulness of the method at broader scale, SSU rRNA molecules from representative organisms of the archaeal domain of life were analyzed (Fig. 4). Shape trees grouped archaeal rRNA structures in a monophyletic group (78% BS), in which the Crenarchaeota and Euryarchaeota kingdoms were clearly identified. Topologies were only moderately supported but resembled those inferred from rRNA se-

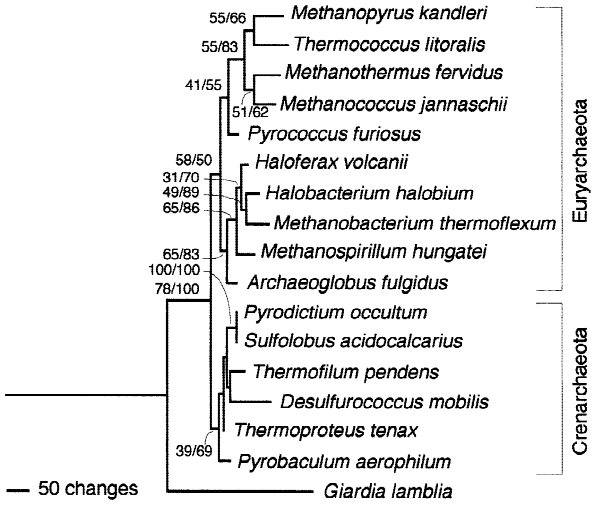

Fig. 4. Phylogenetic reconstruction of an archaeal tree based on the secondary structure of complete SSU rRNA sequences. A total of 197 informative out of 361 characters were analyzed. Two trees (1566 steps; CI = 0.612, RI = 0.453; $g_1 = -0.566$; PTP test, $p = 0.001$) were retained after a heuristic search with TBR branch swapping (one of them shown). Characters were also reweighted by maximum value of rescaled consistency indices in three successive iterations. A single tree (338 steps; CI = 0.730, RI = 0.693; $g_1 = -1.861$; PTP test, $p = 0.001$) was obtained. All trees shared similar topologies. BS values >50% for unweighted and weighted analysis are shown above nodes. The topologies recovered in shape and sequence trees were congruent (PD = 14, SD = 0.093; $p < 0.01$).

quence (Woese et al. 1990; Barns et al 1996; Tourasse and Gouy 1999). Congruence between shape and sequence trees could not be rejected when testing for homogeneity of data partitions ($p = 0.620$). The exclusion of the eukaryotic outgroup (*Giardia lamblia*) from analysis did not alter the root or topology of the archaeal tree (data not shown). These results, and other similarly obtained from representative taxa in Bacteria and Eucarya (G. Caetano-Anollés, mns. in preparation), validate character polarization and the rooting approach utilized and show that the secondary structure of rRNA contains phylogenetic signal capable of grouping organisms of highly divergent lineage. This prompted a study of the origin and diversification of rRNA structure encompassing all primary organismal domains.

Global Rooting of the Tree of Life

Universal phylogenetic trees were inferred from the secondary structure of the small and large subunits of rRNA (Figs. 5 and 6). Trees branched in three major monophyletic groups corresponding to the three domains of life. These groups were supported by 59–98% and 91–99% BS values for SSU and LSU rRNA, respectively. The monophyly of Archaea, Bacteria, and Eucarya parallels that observed in rRNA comparative sequence analysis (Woese 1987; Gouy and Li 1989; Woese et al. 1990; De Rijk et al. 1995), and suggest that diversity originated in three initial dramatic evolutionary events. While the monophyly of individual domains was essentially noncontroversial [with an exception in the Archaea (Tourasse and Gouy 1999)], the topology of the trees was striking. Phylogenies were rooted in the eukaryotic branch, the sisterhood of Archaea and Bacteria being supported by 50% and 97% BS values for SSU and LSU rRNA, respectively. The rooting suggests that it is equally parsimonious to assume that either ancestral unicellular eukaryotes or prokaryotes gave rise to all extant life forms.

Ancestral molecules inferred from structural phylogenetic analysis were used to visualize which structural features in rRNA molecules had changed in the course of evolution. For example, the ancestral LSU molecule contains features that are absent or considerably reduced in Archaea, Bacteria, plastids, and mitochondria (e.g., areas G5_n and H1_n) and some eukaryotes (e.g., C1_n, D4_1 and E20_n) (Fig. 2). These features generally coincide with the hypervariable expansion segments [D(ivergent) domains] responsible for the large size of eukaryotic LSU rRNA (Hassouna et al. 1984). The shape trees are therefore consistent with an overall trend towards molecular simplification, especially in the prokaryotic domains of life.

Results challenge the prokaryotic dogma that states that the last universal common ancestor [the "cenancestor" (Fitch and Upper 1987)] was bacterial-like in genomic and cellular organization, and supports an eukaryotic ancestry (Reanney 1974; Doolittle 1978; Darnell 1978) and the hypothesis that many prokaryotic features originated by simplification through gene loss and nonorthologous displacement (Forterre and Philippe 1999). Note that an eukaryotic rooting has been proposed for other molecules [e.g. SRP (Brinkmann and Philippe 1999)] and provides a simple explanation for the close genomic relationship of Archaea and Bacteria (Koonin et al. 1997).

The trees reconstructed from SSU rRNA structure depicted the unprecedented diversity of rRNA sequence observed in Eucarya, with protoctist lineages being more diverse than those of Archaea and Bacteria put together (Fig. 5). This was not evident in LSU rRNA shape trees (Fig. 6). The SSU rRNA grouping of eukaryal lineages matched for the most part those inferred by comparative sequence analysis and traditional classification. Shape trees continued to show amitochondriate Archezoa (microsporidia, diplomonads, and trichomonads) branching before unicellular eukaryotes with functional mitochondria. This topology is highly debated (Embley and Hirt 1998) and the placement of fast evolving lineages deemed impossible due to GC content and rate heterogeneity in rRNA sequence (Philippe and Germot 2000). Shape phylogenies also exhibited discordance. This usually coincided with poorly supported branches, such as the polyphyly of animals in LSU rRNA trees. In other

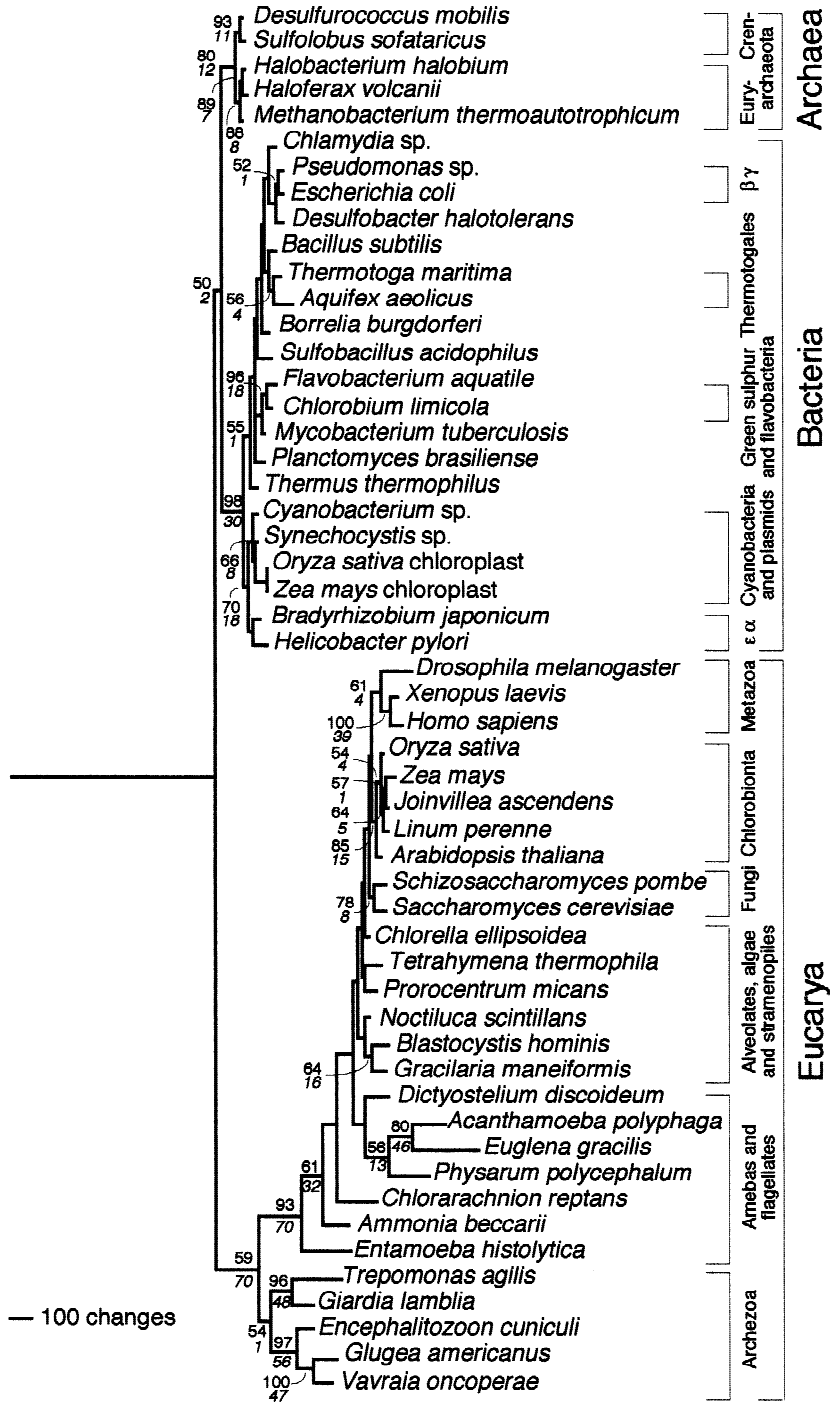

Fig. 5. Phylogenetic reconstruction of a universal tree based on the secondary structure of SSU rRNA. A total of 460 informative out of 662 characters were analyzed. Two trees (7193 steps; CI = 0.384, RI = 0.770; $g_1 = -0.270$; PTP test, $p = 0.001$) were retained after a heuristic search with TBR branch swapping. The rooted shape tree shown is congruent with the 50% majority-rule consensus. Successive character weighting resulted in a tree of 1662 steps (CI = 0.473, RI = 0.866; $g_1 = -0.305$; PTP test, $p = 0.001$). Trees exhibit similar topologies. BS values >50% and decay indexes are shown above and below nodes.

instances, major incongruences were very well supported, such as the grouping of fungi and plants or mitochondria and chloroplasts in LSU rRNA trees. It is important to recognize that shape phylogenies can also be sensitive to tree reconstruction artifacts that are characteristic of sequence analysis. Factors that can incorporate bias in the analysis could include mutational saturation, variation of evolutionary rates across sites, and covarion structure (cf. Philippe and Forterre 1999; Philippe and Germot 2000). Future phylogenetic reconstruction studies should be fine-tuned by selective weighting of characters according to their relative contribution to molecular stability and the relative rates of substitution in their encoding sequences.

Evolution of Ribosomal tRNA Binding Domains

The evolution of protein biosynthesis defines the origins of modern biochemistry, and can explain how ancestral life of a proposed RNA world (Gilbert 1996; Joyce 1991) gave rise to modern organisms (Reanney 1974; Doolittle

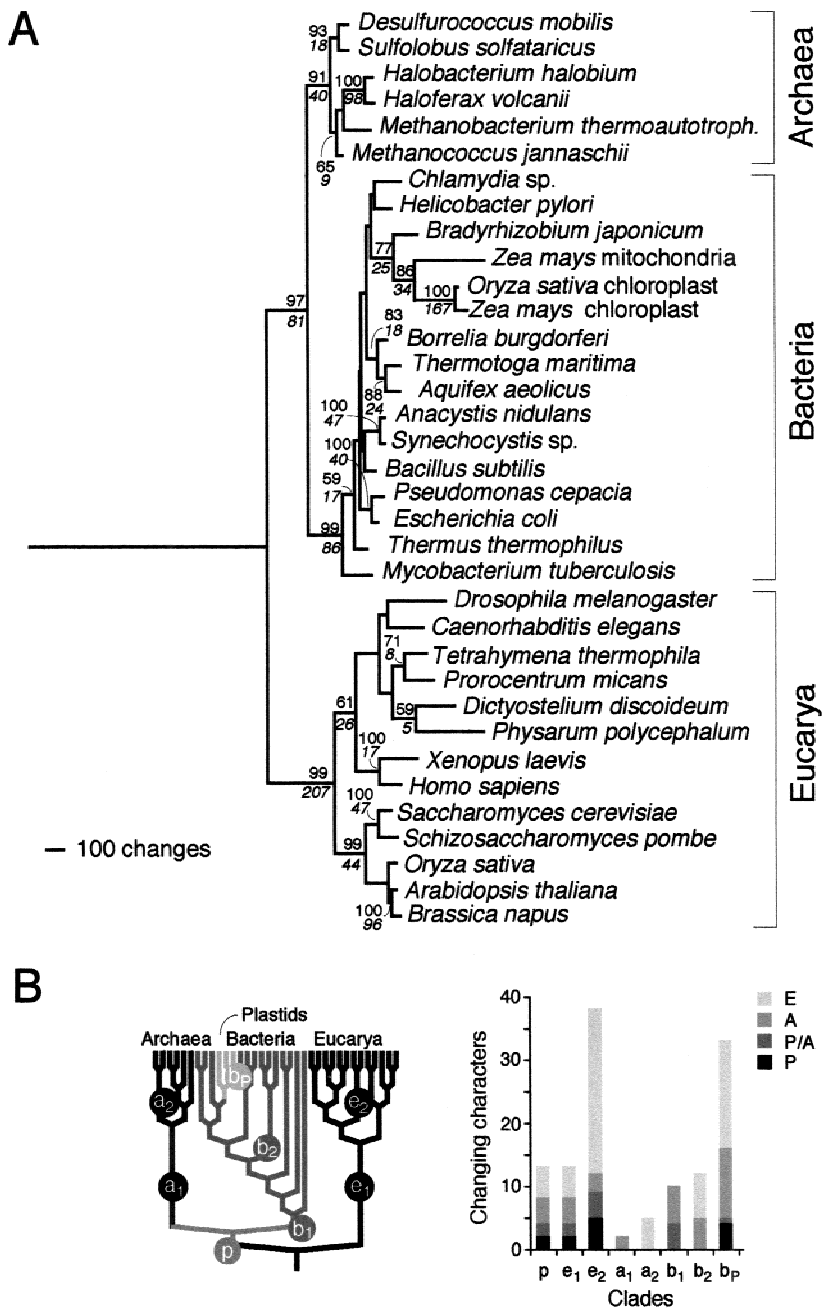

Fig. 6. Phylogenetic reconstruction of a universal tree based on the secondary structure of LSU rRNA. **(A)** Single most-parsimonious tree (9371 steps; CI = 0.447, RI = 0.719; $g_1 = -0.420$; PTP test, $p = 0.001$) retained after a heuristic search with TBR branch swapping. A total of 650 informative out of 848 structural characters were analyzed. BS values >50% and decay indexes are shown above and below nodes. **(B)** The evolution of structural characters defining tRNA-binding domains was traced on the LSU rRNA shape tree. Stacked bar charts show structural characters changing in P, A, shared P and A, and E binding sites. Changes correspond to major clades of the LSU universal tree.

1978; Darnell 1978; Poole et al. 1998; Forterre and Philippe 1999). If RNA predates protein as a molecular catalyst, pivotal RNA molecules such as rRNA or tRNA can be considered relics of the RNA world and extant indicators of the course of evolution in early life. Based on inferred molecular fossils (Jeffares et al. 1998), a model was proposed in which the cenancestor had an eukaryotic-like architecture (Poole et al. 1998). In this model, the proto-ribosome precursor was an ancient RNA replicase that was recruited into the role of protein synthesis by processing amino acid-tagged tRNA substrates. Since naked RNA containing the peptydil transferase center in LSU rRNA can catalyze the formation of peptide bonds (Nitta et al. 1998), it is safe to assume that an ancestral LSU rRNA was a major component of the tRNA-processing proto-ribosome. In order to gain a novel insight into the evolution of protein biosynthesis, the origin and diversification of tRNA binding domains (A, E, and P sites) involved in the translational cycle (Table 1) were traced in LSU rRNA molecules (Fig. 6B). Phylogenetic analysis revealed that binding sites were congruent with other LSU characters ($p = 0.477$) and produced a tree rooted in the Eucarya with monophyletic groups corresponding to the three domains of life. Therefore, tRNA binding sites appear to have evolved concurrently with the rest of the LSU rRNA molecule. Character reconstruction analysis showed that most changes in the P site were ancestral (with an exception in chloroplasts) and

most in A and E sites were derived, suggesting a more primitive nature of the P binding domain. An unusually large number of structural characters changed in plastids. This could be attributed, for example, to accelerated change triggered by endosymbiosis. Clearly, the tracing of structural characters here described provides a new tool to study macromolecular evolution at the phylogenetic level.

Secondary Structure, Evolution, and Phylogenetic Inference

The evolutionary role of RNA structure here uncovered is of paramount importance. The phylogenetic tracing of structural transformation shows that RNA molecules evolve to attain high conformational order [much of which appears intrinsic (Higgs 1994; Schultes et al. 1999)]. This search for greater structural stability and uniqueness imposes a directionality principle on secondary structure that increases thermodynamic entropy, is entirely based on the statistical mechanics of molecules, and could be a major factor constraining the sequence of coding DNA. A molecular tendency towards order and stability supports a thermodynamic theory of evolution (Gladyshev 1978; Black 1978; Gladyshev and Ershov 1982) in which free-energy dissipation occurs at the expense of increasing the chaos of the rest of the universe, and is optimized as biological systems become more refined or complex.

The tendency towards order helps move structures in evolutionary time through adaptive landscapes. This movement is not completely free. Instead, it is constrained by biological function and the mapping of sequence into structure, making some structures distant or completely inaccessible (Fontana and Schuster 1998b). In this regard, the evolutionary tracing of structural characters could reveal some of the topography of this complex and flowing landscape.

Structural analysis at various taxonomical levels shows that natural selection occurs early in the information flow that originates in nucleic acid sequence and ends in an adapted phenotype. Congruent hierarchical patterns of variation in primary and secondary structure suggest a direct evolutionary link between genotype (sequence) and phenotype (structure) at the molecular level. This link is not greatly constrained by the function of the evolved RNA species analyzed, and could therefore operate at levels other than the organismal unit. Most informative characters are variable structural features that generally portray regions other than those functionally conserved (defined by positional covariance) or involved in known catalytic or ligand binding functions. The few characters that may not fully comply with assumptions of state transformation (e.g., unpaired regions constrained by unknown pseudoknot interactions) will tend to be invariable and should introduce little bias in the analysis. Therefore, shape and sequence appear to evolve concordantly and independently from adaptations to specific functions and molecular environments. However, there are no known physical processes capable of modifying shape directly and inheritably besides those driven by fitness, mutation, and evolutionary constraint on underlying sequence. Schultes et al. (1999) recently proposed the organization of RNA sequence space in nested subsets of increasing biological relevance, with function arising in the course of evolution from specific adaptations in structure and these structural phenotypes from sequences encoding molecules with well-ordered conformations. The proposed link between structure and sequence here uncovered is consistent with this hypothesis but requires of an overlapping structural diversification mechanism that is concurrent with functional adaptation (evolutionary design) and self-organization (order) and can fully explain the origin of structure and function in RNA.

The cladistic method here proposed constitutes a new tool for phylogenetic inference that complements classical methods of primary sequence comparison. Both, sequence and structural analysis can reconstruct the evolutionary history of an individual molecule and they can do this congruently. Consequently, shape phylogenies can be used to confirm results obtained by classical phylogenetic analysis or complement the comparative analysis of sequences that contain many indels or are difficult to align. The method here proposed is at present inefficient, sometimes failing to reconstruct indisputable clades. This limits its widespread use in the study of molecular evolution, pending a more detailed analysis of its true potential. At this time it is difficult to evaluate the comparative performance of one or the other method. Evolution is expected to be traced differently in each, as characters (nucleotides and structural features) are the subject of different mutation and selection mechanisms and carry their own and distinct phylogenetic signal. However, one unique feature of the structure-based phylogenetic method here proposed is its ability to produce rooted topologies capable of establishing direction of evolutionary change at the molecular level. This feature can be very useful in applications where suitable outgroups or paralogous sequences are not available.

Acknowledgments I gratefully acknowledge G. E. Caetano-Anollés for support and help in character coding, A. Holst-Jensen, W.D. Bauer, and J. Felsenstein for valuable comments, and V. Knudsen for computer programs. A preliminary account of parts of this study was presented at the Ninth International Congress on Molecular Plant-Microbe Interactions, Amsterdam, The Netherlands, on July 25–30, 1999.

References

Asai T, Zaporojets D, Squires C, Squires CL (1999) An *Escherichia coli* strain with all chromosomal rRNA operons inactivated: complete exchange of rRNA genes between bacteria. Proc Natl Acad Sci USA 96:1971–1976

Ban N, Nissen P, Hansen J, Moore PB, Steitz TA (2000) The complete atomic structure of the large ribosomal subunit at 2.4 Å resolution. Science 289:905–920

Barns SM, Delwiche CF, Palmer JD, Pace NR (1996) Perspectives on archaeal diversity, thermophily and monophyly from environmental rRNA sequences. Proc Natl Acad Sci USA 93:9188–9193

Black S (1978) On the thermodynamics of evolution. Persp Biol Med 21:348–356

Boysen M, Borja M, del Moral C, Salazar O, Rubio V (1996) Identification at strain level of *Rhizoctonia solani* AG4 isolates by direct sequence of asymmetric PCR products of the ITS regions. Curr Genet 29:174–181

Brinkmann H, Philippe H (1999) Archaea sister group of Bacteria? Indications from tree reconstruction artifacts in ancient phylogenies. Mol Biol Evol 16:817–825

Bull J, Wichman H (1998) A revolution in evolution. Science 281:1959

Caetano-Anollés G (2000) Evolution of RNA secondary structure and genome size in the grasses. Plant Mol Biol Rep 18:S21–3

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

Cate JH, Yusupov MM, Yusupova GZ, Earnest TN, Noller HF (1999) X-ray crystal structure of 70S ribosome functional complexes. Science 285:2095–2104

Darnell Jr JE (1978) On the origin of prokaryotes. Science 202:1257–1260

De Rijk P, Van de Peer Y, Van den Broeck I, De Wachter R (1995) Evolution according to large ribosomal subunit RNA. J Mol Evol 41:366–375

Doolittle WF (1978) Gene in pieces: were they ever together. Nature 272:581–582

Doolittle WD (1999) Phylogenetic classification and the universal tree. Science 284:2124–2128

Draper DE (1996) Strategies for RNA folding. Trends Biochem Sci 21:145–149

Embley TM, Hirt RP (1998) Early branching eukaryotes? Curr Opinion Gen Dev 8:624–629

Farris JS, Kållersjö M, Kluge AG, Bult C (1995) Testing significance of incongruence. Cladistics 10:315–319

Fitch WM, Upper K (1987) The phylogeny of tRNA sequences provides evidence for ambiguity reduction in the origin of the genetic code. Cold Spring Harbor Symp Quant Biol 52:759–767

Fontana W, Schuster P (1998a) Continuity in evolution: on the nature of transitions. Science 280:1451–1455

Fontana W, Schuster P (1998b) Shaping space: the possible and the attainable in RNA genotype-phenotype mapping. J Theor Biol 194: 491–515

Fontana W, Konings DA, Stadler PF, Schuster P (1993) Statistics of RNA secondary structures. Biopolymers 33:1389–1404

Forterre P, Philippe H (1999) Where is the root of the universal tree of life? BioEssays 21:871–879

Fox DL, Fisher DC, Leighton LR (1999) Reconstructing phylogeny with and without temporal data. Science 284:1816–1819

Gilbert W (1986) The RNA world. Nature 319:618

Gladyshev GP (1978) On the thermodynamics of biological evolution. J Theor Biol 75:425–441

Gladyshev GP, Ershov YA (1982) Principles of the thermodynamics of biological systems. J Theor Biol 94:301–343

Gogarten JP, Kibak H, Dittrich P, Taiz L, Bowman EJ, Bowman BJ, Manolson MF, Poole RJ, Date T, Oshima T, Konishi I, Denda K, Yoshida M (1989) Evolution of the vacuolar H^+-ATPase: implications for the origin of eukaryotes. Proc Natl Acad Sci USA 86: 6661–6665

Gouy M, Li W-H (1989) Phylogenetic analysis based on rRNA sequences supports the archaebacterial rather than the eocyte tree. Nature 339:145–147

Green R, Noller HF (1997) Ribosomes and translation. Annu Rev Biochem 66:679–716

Gutell RR, Larsen N, Woese CR (1994) Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. Microbiol Rev 58:10–26

Hassouna N, Michot B, Bachellerie JP (1984) The complete nucleotide sequence of mouse 28S rRNA gene. Implications for the process of size increase of the large subunit rRNA in higher eukaryotes. Nucleic Acids Res 12:3563–3583

Higgs PG (1993) RNA secondary structure: a comparison of real and random sequences. J Phys I France 3:43–59

Hsiao C, Chatterton NJ, Asay KH, Jensen KB (1994) Phylogenetic relationships of 10 grass species: an assessment of phylogenetic utility of the internal transcribed spacer region in nuclear ribosomal DNA in monocots. Genome 37:112–120

Huynen MA, Hogeweg P (1994) Pattern generation in molecular evolution: exploitation of the variation in RNA landscapes. J Mol Evol 39:71–79

Huynen MA, Konings DAM, Hogeweg P (1993) Multiple coding and the evolutionary properties of RNA secondary structure. J Theor Biol 165:251–267

Huynen MA, Stadler PF, Fontana W (1996) Smoothness with ruggedness: the role of neutrality in adaptation. Proc Natl Acad Sci USA 93:397–401

Iwabe N, Kuma K, Hasegawa M, Osawa S, Miyata T (1989) Evolutionary relationship of archaebacteria, eubacteria, and eukaryotes inferred from phylogenetic trees of duplicated genes. Proc Natl Acad Sci USA 86:9355–9359

James BD, Olsen GJ, Pace NR (1989) Phylogenetic comparative analysis of RNA secondary structure. Meth Enzymol 180:227–239

Jeffares DC, Poole AM, Penny D (1998) Relics from the RNA world. J Mol Evol 46:18–36

Joyce GF (1989) RNA evolution and the origins of life. Nature 338: 217–224

Kauffman SA (1993) The origins of order. Oxford Univ. Press, New York

Kauffman SA, Johnsen S (1991) Coevolution to the edge of chaos: coupled fitness landscapes, poised states, and coevolutionary avalanches. J Theor Biol 149:467

Kollipara KP, Singh RJ, Hymowitz T (1997) Phylogenetic relationships in the genus *Glycine* Willd based on sequences from the ITS region of nuclear rDNA. Genome 40:57–68

Koonin EV, Mushegian AR, Galperin MY, Walker DR (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the Archaea. Mol Microbiol 25:619–637

Lewontin RC (1974) The genetic basis of evolutionary change. Columbia Univ. Press, New York

Maddison WP, Maddison DR (1999) MacClade: Analysis of phylogeny and character evolution, version 3.08. Sinauer, Sunderland, MA

Mathews DW, 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

Neff N (1986) A rational basis for a priori character weighting. Syst Zool 35:110–123

Nelson GJ (1973) The higher-level phylogeny of the vertebrates. Syst Zool 22:87–91

Nitta I, Kamada Y, Noda H, Ueda T, Watanabe K (1998) Reconstitution of peptide bond formation with *Escherichia coli* 23S ribosomal RNA domains. Science 281:666–669

Page RDM (1993) COMPONENT, tree comparison software for Microsoft Windows, v. 2.0. The Natural History Museum, London

Philippe H, Forterre P (1999) The rooting of the universal tree is not reliable. J Mol Evol 49:509–523

Philippe H, Germot A (2000) Phylogeny of eukaryotes based on ribosomal RNA: long-branch attraction and models of sequence evolution. Mol Biol Evol 17:830–834

Poole AM, Jeffares DC, Penny D (1998) The path from the RNA world. J Mol Evol 46:1–17

Reanney DC (1974) On the origin of prokaryotes. Theor Biol 48:243–251

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 structures. Proc R Soc Lond 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

Swofford DL (1999) Phylogenetic analysis using parsimony and other programs (PAUP*), version 4. Sinauer, 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

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins H (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25:4876–4882

Tourasse NJ, Gouy M (1999) Accounting for evolutionary rate variation among sequence sites consistently changes universal phylogenies deduced from rRNA and protein-coding genes. Mol Phylogenet Evol 13:159–168

Van de Peer Y, De Rijk P, Wuyts J, Winkelmans T, De Watcher R (2000) The European small subunit ribosomal RNA database. Nucleic Acids Res 28:175–176

Woese CR (1987) Bacterial evolution. Microbiol Rev 51:221–271

Woese C (1998) The universal ancestor. Proc Natl Acad Sci USA 95:6854–6859

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria and Eucarya. Proc Natl Acad Sci USA 87:4576–4579

Wright SL (1932) The roles of mutation, inbreeding, crossbreeding and selection in evolution. Proc 6th Intl Congress Genetics 1:356–366

Wuyts J, De Rijk P, Van de Peer Y, Winkelmans T, De Watcher R (2001) The European large subunit ribosomal RNA database. Nucleic Acids Res 29:175–177

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

Zuker M (1989) On finding all suboptimal foldings of an RNA molecule. Science 244:48–52