J Mol Evol (2005) 61:524–530
DOI: 10.1007/s00239-004-0315-1

JOURNAL OF MOLECULAR EVOLUTION

tRNA Creation by Hairpin Duplication

Jeremy Widmann,[1] Massimo Di Giulio,[2] Michael Yarus,[3] Rob Knight[1]

[1] Department of Chemistry and Biochemistry, University of Colorado, Boulder, CO 80309, USA
[2] International Institute of Genetics and Biophysics, CNR, Via G. Marconi 10, Naples 80125, Italy
[3] Department of Molecular, Cellular and Developmental Biology, University of Colorado, Boulder, CO 80309, USA

Received: 4 November 2004 / Accepted: 3 May 2005 [*Reviewing Editor:* Dr. Niles Lehman]

Abstract. Many studies have suggested that the modern cloverleaf structure of tRNA may have arisen through duplication of a primordial hairpin, but the timing of this duplication event has been unclear. Here we measure the level of sequence identity between the two halves of each of a large sample of tRNAs and compare this level to that of chimeric tRNAs constructed either within or between groups defined by phylogeny and/or specificity. We find that actual tRNAs have significantly more matches between the two halves than do random sequences that can form the tRNA structure, but there is no difference in the average level of matching between the two halves of an individual tRNA and the average level of matching between the two halves of the chimeric tRNAs in any of the sets we constructed. These results support the hypothesis that the modern tRNA cloverleaf arose from a single hairpin duplication prior to the divergence of modern tRNA specificities and the three domains of life.

Key words: tRNA — Hairpin duplication — Cloverleaf

Introduction

Many lines of evidence suggest that the two halves of tRNA may be evolutionarily distinct. For example, the "operational code" that links amino acids to tRNAs depends only on the acceptor stem for certain amino acids (Schimmel and Henderson 1994), and aminoacyl tRNA synthetases can even charge minihelices that resemble only one half of the tRNA molecule (Tamura and Schimmel 2001). These charged minihelix structures have been shown to function in peptide synthesis and may have been part of the primordial protein synthesis machinery (Dick and Shamel 1995). It has also been suggested that the top half of modern tRNAs has an ancient origin in replication and is recognized separately by RNaseP, the CCA-adding enzyme, telomerase, and aminoacyl-tRNA synthetases (Weiner and Maizels 1987; Maizels and Weiner 1994). The 3′ half of modern tRNAs has been proposed to be older than the 5′ half due to its base composition and repetitive sequence patterns (Eigen and Winkler-Oswatitsch 1981). More recently, it has been shown that the archaeon *Nanoarchaeum equitans* can create functional tRNAs from the 3′ and 5′ tRNA halves, which are encoded by different loci and trans-spliced to form the final product (Randau et. al. 2005).

Similarities between nucleotides at comparable positions within the two halves of the tRNA molecule have often been taken as evidence that the modern cloverleaf structure arose through direct duplication of a hairpin (Jukes 1995; Di Giulio 1995 and references cited therein). If this duplication theory is correct, corresponding positions in the two halves of each modern tRNA molecule should match more than chance predicts (i.e., should have greater sequence identity). Additionally, the halves of *different* tRNA molecules should match to greater or lesser extents depending on how many tRNA-creating

Correspondence to: Rob Knight; *email:* rob@spot.colorado.edu

duplication events occurred. Specifically, tRNA halves that came from the same duplication event should match each other better than tRNA halves that came from different duplication events, even if these halves are not found in the same modern tRNA molecule.

There are several possibilities for the number and timing of these duplication events relative to the divergence of the different amino acid specificities, and of the three domains of life (eukaryotes, bacteria, and archaea). All modern tRNAs stemming from a particular duplication event should have about the same number of matches between the two tRNA halves, even if one half comes from one tRNA and the other half comes from another tRNA. This equivalent level of similarity stems from the fact that the duplication creates two identical halves, but in the absence of convergent evolution or recombination, all tRNAs are expected to diverge in sequence equally after the duplication event. Conversely, two tRNAs that do not stem from the same duplication event should have fewer matches between the first half of one tRNA and the second half of the second tRNA, because the duplication event would replicate changes that occurred in one of the original hairpins but not the other. We emphasize that the base pairs in the cloverleaf structure need not be the same as the base pairs in the original hairpins. For the cloverleaf structure to be more stable than those in the hairpin, either different base pairs must form in the cloverleaf or the two hairpins must not be identical and must originally be partially mismatched. For example, in the modern tRNA structure, the identities between hairpin halves identified by sequence alignment generally do not have the same base pairs as those in the cloverleaf, so that if A pairs with A′ and B pairs with B′ in the original hairpins, A need not pair with B′ and A′ with B in the cloverleaf (Di Giulio 1995). Although sequences that can support both sets of pairing constraints are relatively rare, they are expected to be found at the appreciable frequencies of 1 in ~30 million random sequences (Nagaswamy and Fox 2003) and are thus easily accessible to evolution.

We consider five distinct scenarios for the evolution of modern tRNAs through duplication and divergence (Fig. 1). First, the similarities between the halves of each molecule might be caused by modern selection for function in each tRNA, eliminating the need for any duplication events to explain the similarities. In this scenario, the two halves of a given tRNA should match each other, but would not be expected to match the corresponding halves of another arbitrarily chosen tRNA. Second, the cloverleaf might have arisen once, before either the domains or the amino acid specificities emerged, thus explaining the similarities with a single duplication event (Di Giulio 1992). In this scenario, matching between the two tRNA halves in modern sequences arises through common descent from an ancestral sequence in which the halves matched. Thus, the halves of any two tRNAs would match to about the same degree. Third, the cloverleaf might have arisen independently in each of the three domains by duplication and then diverged to produce different specificities, requiring three duplication events. In this scenario, the tRNAs in each domain arise from an independent duplication, so the halves of two tRNAs from within a single domain should match better than the halves of two tRNAs from different domains. This third scenario has been proposed as a way to explain incongruence in tRNA phylogenies (Di Giulio 1999). Fourth, each tRNA specificity might have arisen through an independent duplication before the three domains diverged, requiring at least 20 duplication events. In this scenario, the tRNAs in each specificity arise from an independent duplication, so the halves of two tRNAs from within a single specificity should match better than the halves of two tRNAs from different specificities. Finally, each specificity in each domain might have arisen from an independent duplication, requiring the largest number of independent duplication events. In this scenario, the tRNAs in each domain and specificity arise from an independent duplication, so the halves of two tRNAs from within a single domain and specificity should match better than the halves of two tRNAs from a different domain and specificity.

Here, we test these scenarios by counting the number of matches between the two halves of chimeric tRNA molecules, where the first half and the second half may either be restricted to come from tRNAs in the same group (by domain, specificity, or both) or be unrestricted. We expected to be able to identify duplication events by finding group restrictions such that tRNA halves from random pairs of tRNAs within a single group have more matches on average than tRNA halves from random pairs of tRNAs from different groups. For example, if tRNAs evolved from separate hairpin duplications for each specificity, we would expect that chimeric tRNAs made from halves of tRNAs with the same specificity would have more matches than chimeric tRNAs made from halves of tRNAs with different specificities.

Methods

To establish that the hairpin duplication scenario was plausible, we first tested whether the similarity between the two halves of real tRNAs was in fact greater than that of random sequences that could fold into the canonical cloverleaf structure. We obtained 5950 sequences from the Sprinzl Genomic tRNA database (Sprinzl and Vassilenko 2003). Starting with previously published alignments and secondary structures of reconstructed ancestral tRNA

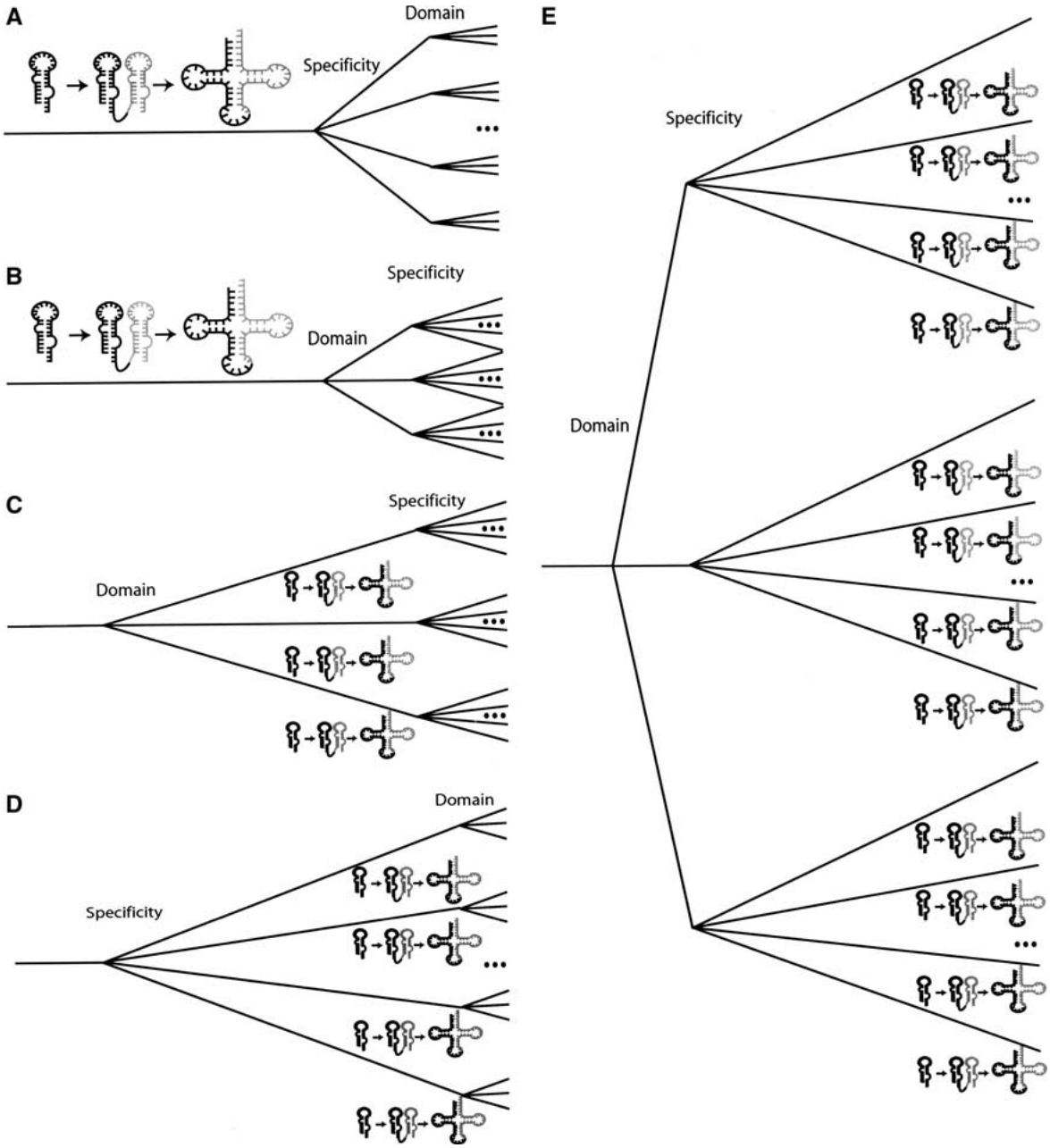

Fig. 1. Five scenarios for duplication of a hairpin to create the modern cloverleaf structure. **a** After an initial duplication, tRNAs diverged into specificities and then into domains. **b** After an initial duplication, tRNAs diverged into domains and then into specificities. **c** Hairpin pre-tRNAs diverged into domains, duplicated in each domain to form cloverleaves, and then diverged into specificities. **d** Hairpin pre-tRNAs diverged into specificities, duplicated in each specificity to form cloverleaves, and then diverged into domains. **e** Hairpin pre-tRNAs duplicated independently in each domain and specificity. See text for discussion.

sequences (Di Giulio 1995), we were able to define the two halves of a tRNA molecule and the specific positions that should match between the two halves. We compared the distribution of matches for real tRNA sequences with that of chimeric tRNA sequences made of halves from within or between specific groups of tRNAs.

tRNA Alignment

The published alignment of reconstructed ancestral sequences was generated with the ALIGN program, which uses the Needleman–Wunsch global alignment algorithm. The alignment of the consensus sequences was generated by inserting gaps at the same positions in the sequences as those in the published alignment. Figure 2 shows the consensus alignment and structure, using the consensus sequence of all the tRNAs in the Sprinzl database rather than those of the ancestral sequences as in previous work (Di Giulio 1995).

Generating Random tRNAs

Random tRNA sequences were constructed by randomizing the nucleotide sequences of the real tRNA sequences in the Sprinzl

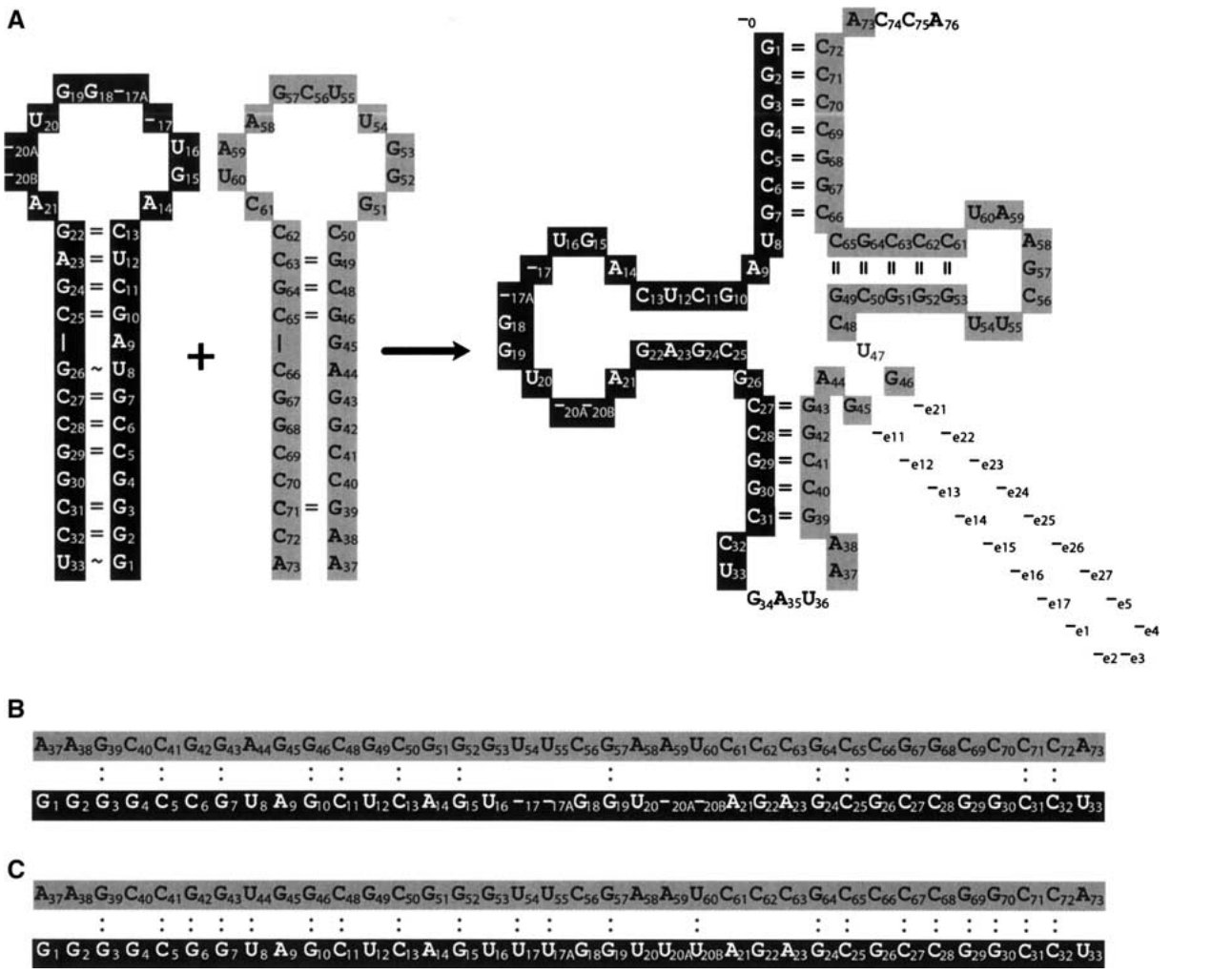

Fig. 2. Matches between the two halves of the modern cloverleaf structure, possibly produced by hairpin duplication. **a** Fusion of two hairpins to form the modern cloverleaf. Bases are numbered as in the Sprinzl database. The most frequent base is shown at each position. **b** Matches between the two halves of the consensus sequence from the Sprinzl database. **c** Matches between the two halves of the reconstructed ancestral tRNA sequence (Di Giulio 1995).

database. For each tRNA sequence in the database, we made one list containing each unpaired base in that tRNA and a second list containing each base pair. Each list was shuffled to randomize the order. This shuffling used the Yates–Fisher algorithm and the Mersenne Twister random number generator as implemented in the Python 2.3 package. We reconstructed the sequence from these two lists so that the structure and base composition were the same as the original tRNA, although the sequence was randomized.

Comparing Matches for Real and Random tRNAs

For each tRNA sequence in the Sprinzl database, we counted the number of times that the corresponding positions in the two halves of the tRNA (as defined in Fig. 2) matched. We repeated this procedure for the set of randomized tRNAs.

Comparing Matches for Real and Chimeric tRNAs

We organized the tRNA sequences in the Sprinzl database into (overlapping) groups as follows: all tRNAs regardless of domain and specificity, all tRNAs in the same domain, all tRNAs with same amino acid specificity, and all tRNAs with the same domain and same specificity. Within each group, we joined the first half of each tRNA to the second half of another, randomly chosen, tRNA. We then counted the number of matches between the first and the second halves of the new, chimeric tRNAs. We compared the distribution of matches from each group of these chimeric tRNAs to that of the actual tRNAs, as identified above.

Results

Real tRNAs have significantly more matches between their two halves than do random tRNAs (Fig. 3) ($t = 15.8$, df $= 11898$, $p = 4.67 \times 10^{-55}$, paired two-sample t-test). The distributions of matches between the two halves of chimeric tRNAs from any combination of domain and/or specificity are essentially identical to the distributions of matches between the halves of real tRNA sequences (Fig. 3).

We also tested whether the specific positions within the tRNA that contributed most to matches

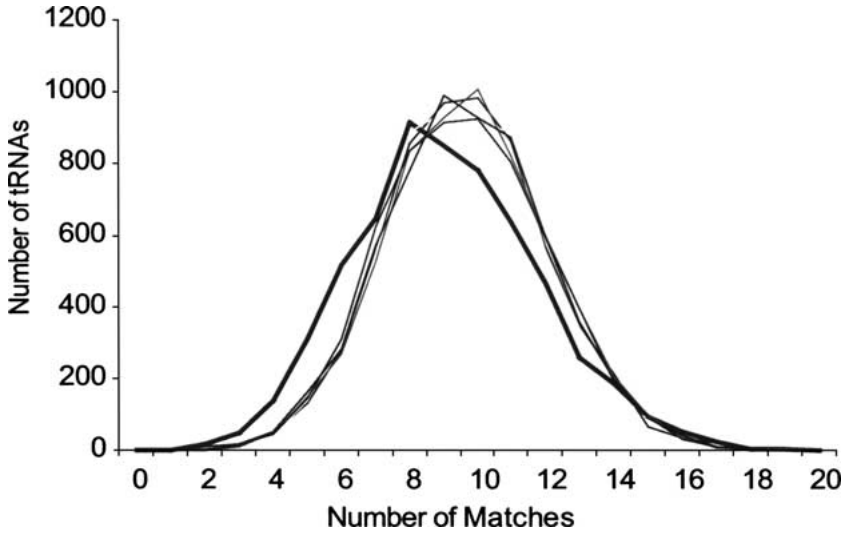

Fig. 3. Distribution of matches between tRNA halves in sequences generated by different models. Individual tRNAs and chimeric tRNAs made by randomly selecting halves from within a domain, a specificity, a domain and specificity, or any two tRNAs (thin lines, statistically indistinguishable from one another) have significantly more matches between the two halves than do random sequences that are generated to allow the base pairs in canonical tRNA structure (thick line). This graph shows the number of matches between the two halves (x axis) plotted against the number of tRNAs with that many matches (y axis).

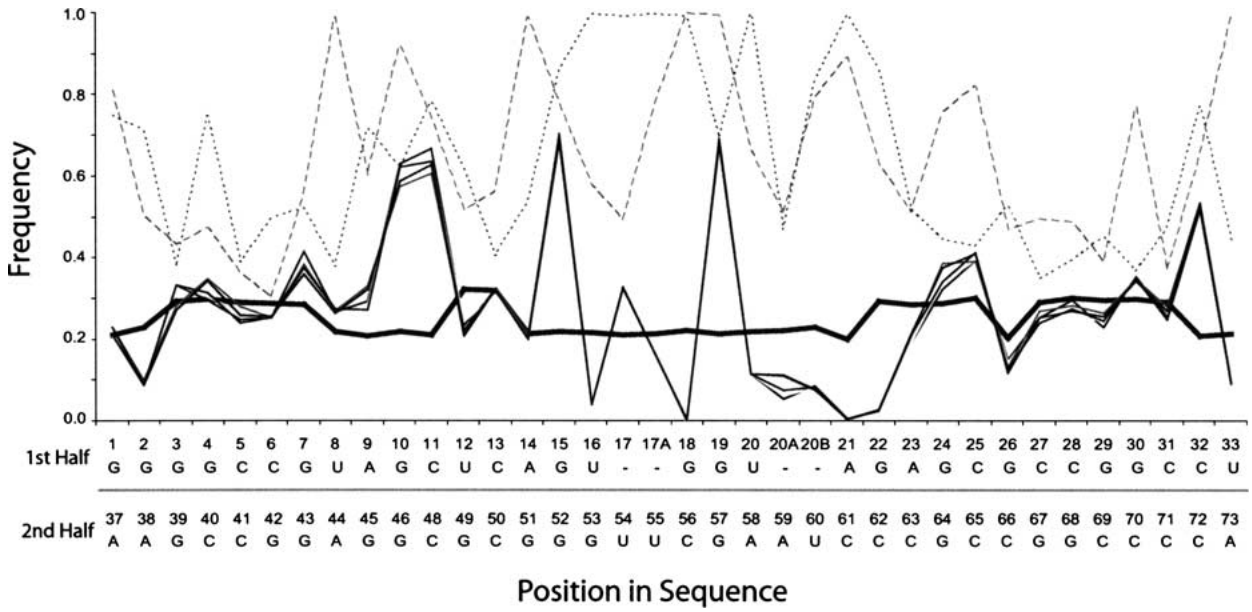

Fig. 4. Frequency (y axis) of matches (solid lines) and conservation of most frequent base (dashed lines) plotted against position within the tRNA sequence (x axis). Match frequencies show the fraction of the time that the two corresponding positions in the first half and the second half are identical. The thick line shows the distribution for randomly generated sequences, while the (very similar) thin lines show the distribution for the actual tRNAs and each of the chimeric sets of tRNAs. Conservation shows the proportion of the most frequent base at each position. The long dashes refer to the conservation in the first half of the sequence, while the short dashes refer to the second half of the sequence. The most frequent base at the position is printed below the position number at the bottom of the graph.

were also highly conserved. The consistent trend, for all chimeric tRNAs, is that several highly conserved positions did contribute to matches (Fig. 4). However, there was no overall correlation between the amount of conservation at a position and the tendency of that position to match between tRNA halves ($p > 0.05$). For comparison, Fig. 5 shows the secondary structure of a tRNA molecule and the calculated percentage conservation of the base at each position based on the sequences from the Sprinzl database.

Discussion and Conclusions

As shown in Figs. 3 and 4, we found no significant difference in the number of matches between the two tRNA halves for any of the chimeric sets we constructed. This observation, combined with the highly significant excess of matches between each of the chimeric sets and the set of random sequences, supports an ancient, monophyletic tRNA origin that predates the divergence of specificities and domains. Consequently, the data support option a or b in

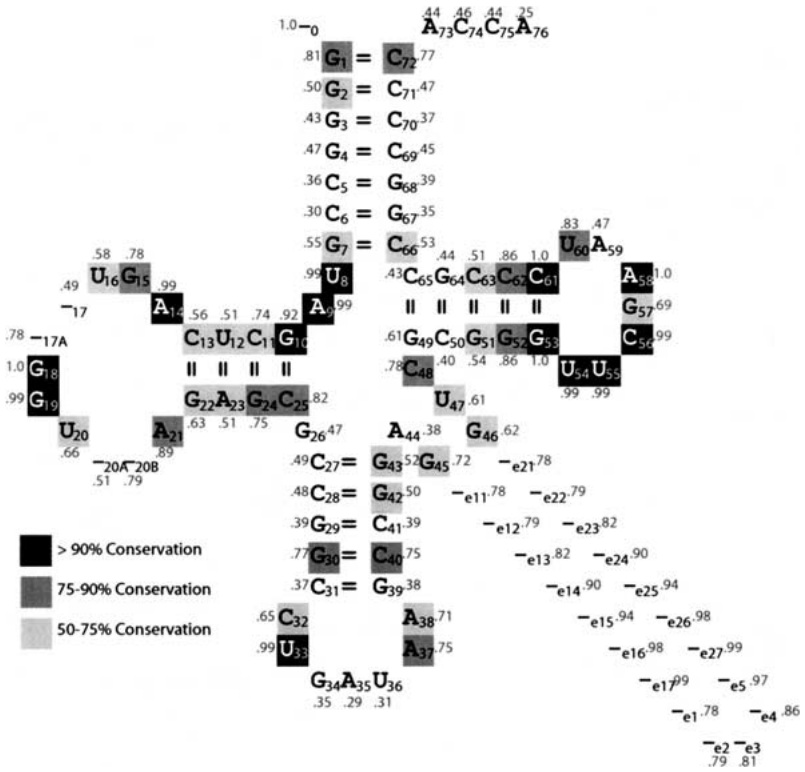

Fig. 5. Canonical cloverleaf structure of tRNA, highlighting conserved positions. Positions with greater than 90% conservation are shown in black, positions with conservation between 75% and 90% are shown in dark gray, and positions with between 50% and 75% conservation are shown in light gray. The most frequent base across all tRNAs is shown, the conventional position number is shown as a black subscript, and the degree of conservation is shown as a gray subscript. Note that positions that are shown as dashes indicate that the most frequent state at that position is for the base to be missing (i.e., a gap).

Fig. 1 equally. Although incongruent tRNA phylogenies have provided much of the evidence for a nonmonophyletic tRNA origin in a duplication, it is often difficult to resolve trees based on sequences of fewer than 500 nucleotides (Nei et al. 1998). tRNA phylogenies can thus be difficult to interpret, since the region conserved across specificities is only 74 nucleotides in length. Switches in tRNA specificity have also been demonstrated by as little as a single nucleotide change (Yaniv et al. 1974; Saks et al. 1998), which might also lead to nonmonophyly of individual specificities even if all tRNAs descended from a single, common ancestor.

Though our outcome is consistent with the idea that tRNAs arose from a single ancestral duplication event and less consistent with duplication on other schedules, some caution is warranted. It is also possible that these data are entirely the result of convergent selection for function. Additionally, it remains conceivable that differences that discriminate among groups have been lost during the vast span of time since the appearance of the modern tRNA repertoire. In this context, it is interesting to note the discrepancy between the largest number of matches between the halves of modern sequences and the number of matches between the halves of the inferred ancestral sequences in Di Giulio (1995). Modern tRNAs average about 9 positions that match between the two halves (Fig. 3), but the ancestral alignment shows 21 matching positions (Di Giulio 1995). It is possible that the ancestral reconstruction technique overcomes some of the loss of information through neutral mutation, although it is also possible that long-branch attraction effects in the parsimony analysis (Felsenstein 1978) lead to difficulties in assigning ancestral states.

If a cloverleaf produced by duplication of an ancient hairpin evolved into modern tRNAs, our results (Figs. 4 and 5) suggest that some primordial similarities between the halves were captured by evolution, particularly in the modern conserved sequences of the D and TΨC stems and loops.

Acknowledgments. This work was supported by a seed grant from the W. M. Keck Foundation RNA Bioinformatics Initiative. We thank members of the Knight and Yarus labs for critical discussion of the manuscript.

References

Di Giulio M (1995) Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule? J Theor Biol 177:95–101

Di Giulio M (1999) The non-monophyletic origin of the tRNA molecule. J Theor Biol 197:403–414

Dick T, Schamel W (1995) Molecular evolution of transfer RNA from two precursor hairpins: implications for the origin of protein synthesis. J Mol Evol 41:1–9

Eigen M, Winkler-Oswatitsch R (1981) Transfer-RNA, an early gene? Naturwissenschaften 68:282–292

Felsenstein J (1978) Cases in which parsimony and compatibility methods will be positively misleading. Syst Zool 27:401–410

Jukes TH (1995) A comparison of mitochondrial tRNAs in five vertebrates. J Mol Evol 40:537–540

Maizels N, Weiner A (1994) Phylogeny from function: Evidence from the molecular fossil record that tRNA originated in replication, not translation. Proc Natl Acad Sci USA 91:6729–6734

Nagaswamy U, Fox GE (2003) RNA ligation and the origin of tRNA. Orig Life Evol Biosph 33(2):199–209

Nei M, Kumar S, Takahashi K (1998) The optimization principle in phylogenetic analysis tends to give incorrect topologies when the number of nucleotides or amino acids used is small. Proc Natl Acad Sci USA 95:12390–12397

Randau L, Münch R, Hohn M, Jahn D, Söll D (2005) Nanoarchaeum equitans creates functional tRNAs from separate genes for their 5′- and 3′-halves. Nature 433:537–541

Saks ME, Sampson JR, Abelson J (1998) Evolution of a transfer RNA gene through a point mutation in the anticodon. Science 279(5357):1665–1670

Schimmel P, Henderson B (1994) Possible role of aminoacyl-RNA complexes in noncoded peptide synthesis and origin of coded synthesis. Proc Natl Acad Sci USA 91(24):11283–11286

Sprinzl M, Vassilenko KS (2005) Compilation of tRNA sequences and sequences of tRNA genes. Nucleic Acids Res 1:33, D139–D140

Tamura K, Schimmel P (2001) Oligonucleotide-directed peptide synthesis in a ribosome- and ribozyme-free system. Proc Natl Acad Sci USA 98:1393–1397

Weiner A, Maizels N (1987) tRNA-like structures tag the 3′ ends of genomic RNA molecules for replication: Implications for the origin of protein synthesis. Proc Natl Acad Sci USA 84:7383–7387

Yaniv M, Folk WR, Berg P, Soll L (1974) A single mutational modification of a tryptophan-specific transfer RNA permits aminoacylation by glutamine and translation of the codon UAG. J Mol Biol 86:245–260