Demonstrating Evolutionary Relationships Between Macromolecular Sequences through Mutual Relationships with a Third Sequence

Mark P. Staves, David P. Bloch

Department of Botany, University of Texas at Austin, Austin, Texas 78713, U.S.A.

and

James C. Lacey, Jr.

Department of Biochemistry, University of Alabama at Birmingham, Birmingham, Alabama 35294, U.S.A.

Z. Naturforsch. **43c,** 476–478 (1988);
received December 30, 1987/February 11, 1988

Euglena Chloroplast, Yeast, rRNA, tRNA, Sequence Comparisons

Corresponding sites of the *Euglena* chloroplast and yeast small subunit ribosomal RNAs (rRNAs) show only an insignificant match with each other but show extensive matches with the *Euglena* chloroplast $tRNA^{arg}$. The match with the tRNA extends farther toward the 5′ end of the *Euglena* rRNA and toward the 3′ end of the yeast rRNA. The expected number of such configurations given the number of RNAs searched is about 1 in 100,000. Comparison of two sequences with a third sequence frequently reveals relationships where pairwise comparisons fail to do so.

Introduction

In previous reports we have described matching sequences between transfer RNAs (tRNAs) and small subunit ribosomal RNAs (rRNAs) [1, 2]. The matches have expected numbers ranging from 0.1 to 10^{-6} and are found in 30–40% of the searches. Comparisons were made using tRNA and small subunit rRNA sequences from highly divergent sources representing archaebacteria, eubacteria, eukaryotes, chloroplasts and mitochondria. Matches are found when searches are conducted using tRNA sequences from one organism and rRNA sequences from another, just as they are when the searches are between RNAs from the same species. Additionally, the matches are found with similar frequencies indicating that the matches reflect true homology (*i.e.* common ancestry) rather than conversion [2].

Simple pairwise sequence comparisons between macromolecules, however, may not reveal relationships which may be revealed by their mutual relationships with a third molecule. Here we describe data for one case, $tRNA^{arg}$ from *Euglena* chloroplasts and the small subunit rRNAs from the same organism and from yeast, and discuss mechanisms for its origins. The phenomenon may be a manifestation of "genes in pieces" [3] as related to structural RNAs.

Reprint requests to M. P. Staves, Department of Biochemistry, University of Alabama at Birmingham, Birmingham, Alabama 35294, U.S.A.

Verlag der Zeitschrift für Naturforschung, D-7400 Tübingen
0341–0382/88/0500–0476 $ 01.30/0

Materials and Methods

The Los Alamos searching routines [4] were used to find matches between the arginine-encoding tRNA from *Euglena* chloroplast and the small subunit rRNAs from this same source and from yeast. Sequence data were obtained from GenBank (Bolt Baranek Inc.) and from Stutz [5]. The tRNA sequence was standardized to the numbering system of Sprinzl and Gauss [6] and small subunit rRNAs were standardized to *E. coli* [7]. Matching regions were plotted, tRNA *vs.* rRNA, as described by Bloch *et al.* [2] and overlapping matches identified from the plots. Values for the probabilities and expected numbers were calculated according to Bloch *et al.* [1] as modified from Goad and Kanehisa [4]. Under the conditions used for the searches, penalties of −1, 2, 3 for matches, replacements, deletions/insertions, respectively, the expected numbers as calculated are a close approximation of the numbers found using scrambled sequences, being underestimates of the latter by a factor of 0.5 at worst (unpublished data).

Results and Discussion

Fig. 1 shows matches between the *Euglena* chloroplast $tRNA^{arg}$ and the small subunit RNAs from *Euglena* chloroplast and yeast. The default cut-off value of 1×10^{-6} for the probability screened out matches with probabilities higher than this value. The modest region of overlap in which the match is shared by both rRNAs would not be sufficient using this criterion to identify the match between the two rRNAs.

The configuration including the three molecules has a very low probability of occurrence by coincidence. The low probability also translates into a low expected number that takes into consideration the number of opportunities for a given match to be found along and among all of the molecules searched.

The positions of the rRNA sequences in Fig. 1 reflect an overlap of homologous regions based upon

```
1171 UGGGCUACA  CACGUG   CUAC  1189      Euglena chloroplast 16S rRNA
     : : :  : : :  :  : : : : : :  : : :                  E = .0560
  16 UGGACUAGAGCACGUG GCUACGAACU  ACGGAGUCAG 51 Euglena tRNA^arg
                 : : : : :  :  : : : : :  :  : :  : : : : : :  : : :      E = .0097
        1447     GCACGCGCGCUAC A CUGACGGAGCCAG 1473  yeast 18S rRNA
```

for triad, $E = 4.8 \times 10^{-9}$
(1230 threesomes represented by 5 rRNAs and 123 tRNAs were compared).

Fig. 1. Matches between rRNAs from *Euglena* chloroplast and yeast, and $tRNA^{arg}$ from *Euglena* chloroplasts. The numbers of the beginning and end bases as well as E, the expected number of matches of the quality observed, based on coincidence, are shown. The overlapping regions between the rRNAs are homologous based on the secondary structures of the molecules.

the secondary structures of the molecules. Thus the tRNA shows a match beginning with the *Euglena* chloroplast 16S rRNA and continues by matching a contiguous portion of the yeast 18S rRNA. The portion of the tRNA sequence involved in the match extends from base 16–51, corresponding to the D-arm through the beginning of the TΨC stem. Therefore it includes both a highly conserved region (D-arm) and a much less rigorously conserved portion of the tRNA molecule (anticodon arm).

Fig. 2 shows a match between the $tRNA^{arg}$ and a consensus sequence drawn from six additional matches involving homologous regions of rRNAs. The consensus base is represented by that single base which comprises a majority of 50% or better among all the bases in a given position covered by the available matches.

Overlapping regions of tRNAs whose extensions match contiguous regions of an rRNA are fairly common. Eight of these and related configurations have been encountered among comparisons of 104 tRNAs with 5 rRNAs. Related configurations include contiguous associations where the pair of tRNA matches against an rRNA abut against one another, interrupted associations that include mismatches between the pairs, while maintaining corresponding distances spanning the matches in both members, and double pairs of matches in which the spatial relationships are maintained while the rRNA is replaced by a pair of rRNAs. In the present example, the pair of matches involves two rRNAs and a single tRNA.

The configuration in Fig. 1 might be explained as the result of recombination or gene conversion. Presumably the evolutionary history of these RNAs included a large number of such events that resulted in the scattering of matching segments, many of which are large enough to be recognized. It is interesting to speculate whether this history was marked by a period of intense recombinational activity, or whether such activity occurred continuously and still occurs. Comparison of configurations among closely and distantly related organisms may provide a clue. The preservation of the segments must reflect constraints on both mutation and recombination, probably because of interlocking functions that would require compensating changes at multiple sites.

These segments are reminiscent of "genes in pieces" [3] although there is no apparent relation here to introns as there is among the coding genes. The most obvious example of shuffling segments is seen in the displaced homologies [8] or "slippage" [7]. For example, 13 tRNAs share common se-

```
1220 UGGGcuagAGCACGUGGCUAC  AACUGacGGAG  cAgugg  1255  consensus rRNA
     : : :  : : : : : : : : : : : : : : : : : :  : : : :  : : : : : :  : : :  : :    E = 8.01 × 10^-10
  16 UGGACUAGAGCACGUGGCUACGAACU  ACGGAGUCAGGGG  54    Euglena tRNA^arg
```

Fig. 2. Match between *Euglena* chloroplast $tRNA^{arg}$ and consensus RNA. The consensus sequence was obtained from nine overlapping matching sequences between tRNAs and rRNAs in this area. Numbers of positions on the consensus sequence correspond to numbers on *E. coli* 16S rRNA used as a standard. The lower case letters indicate no consensus among several different bases represented at those positions. The bases belonging to the rRNA showing the match with the tRNA were used.

quences in positions 11 through 23. An additional tRNA shares the same sequence at position 2 through 14 [2].

The existence of matches among RNAs belonging to these two classes of molecules that serve obviously different functions, their variable positions among molecules within the same class [1, 2], and their occurrence among molecules belonging to widely disparate groups of organisms [2] suggest the matches are relics of fossil sequences which reflect a common origin for these molecules. Further, these observations indicate that many of the events shaping these molecules may have taken place prior to the last common ancestor of all extant species defined by Cairns-Smith [9] as the earliest organism that contained a genetic apparatus as we know it.

Acknowledgements

This work was supported by grants from the Richard Lounsbery Foundation and the RGK foundation.

[1] D. P. Bloch, B. McArthur, R. Widdowson, D. Spector, R. C. Guimaraes, and J. Smith, J. Mol. Evol. **19,** 420–428 (1983).
[2] D. P. Bloch, B. McArthur, and S. Mirrop, BioSystems **17,** 209–225 (1985).
[3] W. Gilbert, Nature **271,** 501 (1978).
[4] W. Goad and M. Kanehisa, Nuc. Acids Res. **10,** 247–263 (1982).
[5] E. Stutz, Bot. Helv. **94,** 145–159 (1984).
[6] M. Sprinzl and D. Gauss, Nuc. Acids Res. **12,** r1–r58 (1984).
[7] C. Zweib, C. Glotz, and R. Brimacombe, Nuc. Acids Res. **9,** 3621–3640 (1981).
[8] D. P. Bloch, B. McArthur, R. C. Guimaraes, J. Smith, M. Reese, and S. Jayaseelan, submitted.
[9] A. G. Cairns-Smith, Cambridge University Press, London 1982.

Verantwortlich für den Inhalt: A. Klemm
Satz und Druck: Allgäuer Zeitungsverlag GmbH, Kempten