RECENT PROGRESS IN THE DETERMINATION OF THE PRIMARY SEQUENCE OF THE 16 S RNA OF *ESCHERICHIA COLI*

C. EHRESMANN, P. STIEGLER, P. CARBON and J. P. EBEL
Institut de Biologie Moleculaire et Cellulaire du CNRS, 15, Rue René Descartes 67084 Strasbourg Cédex, France

Received 26 October 1977

1. Introduction

During the last few years, the essential role of the 16 S RNA, not only in maintaining the structure of the 30 S subunit, but also in actively participating in the ribosomal functions, has been emphasized. The knowledge of the primary structure of the 16 S RNA is essential for understanding the secondary and tertiary organization of this RNA, the topography of the RNA within the ribosomal subunit, the mechanism of RNA–protein interactions and the participation of the RNA in the function of the ribosome. Since the last publication concerning the primary structure of the 16 S ribosomal RNA from *E. coli* in 1975 [1,2], further progress has been made in our laboratory, enabling us to bring about a number of modifications and clarify certain ambiguities. In this paper, we report the new sequence data.

2. Material and methods

The different techniques used are described [2]. We have developed partial hydrolysis conditions for the 30 S subunits prepared in different ways (salt-washed, unwashed, partially unfolded subunits) which will be described elsewhere. In all cases, the 16 S RNA was ^{32}P-labelled, the partial hydrolysis fragments were fractionated and purified by gel electrophoresis techniques [2] and the sequences were determined using the fingerprinting method [3]. In addition, oligonucleotide digestion with U_2 ribonuclease (Sankyo) was carried out according to [4]. Further characterization (alkaline and venom phosphodiesterase hydrolysis) was performed as described [3]. Identification of digestion products was aided by comigration using known sequences as standards.

3. Results

New partial hydrolysis products have been obtained, allowing us to complete, and in some cases modify, sequences within several areas of the molecule. This additional information can be summarized as follows:

3.1. *Section Q'–F*

The T_1 oligonucleotide ACCAAAG (spot 93d in our designation) had not been localized within the 16 S RNA sequence [2]. It has now been characterized as the result of a cistronic heterogeneity in section Q'F. This oligonucleotide is derived from the sequence ACCĞAAAG by the deletion of a G residue. This finding also allows the completion of the sequence of the pancreatic product previously reported as being G(AAAG.AG)GGGGAC (spot 2), to give the sequence GAAAGAGGGGGAC.

3.2. *Section Q*

A partial reorganization has been performed in section Q in which the pancreatic product AGGU is observed, not as previously reported (py)GAGU. Nevertheless the oligonucleotides contained in this part of the molecule still remain unordered, due to the lack of partial digestion products isolated from this region. The pancreatic products corresponding to this region are: AGC, AGGU, GGGGU, AGU, (GU ?).

3.3. *Section C''*

The relative positions of (G)GGCG and (G)GAAUUACUG have been resolved. The completed sequence may now be written (G)GAAUUACUGGGCG.

A partial T_1 hydrolysis product has been isolated containing GGUUAAUCG at the 5'-position. (G)GGU is therefore a partial product from the reported pancreatic RNAase product G(AG.G_3)U (spot 3). This allows the deduction of the partial sequence G(A.G)GGGU for this pancreatic RNAase oligonucleotide.

3.4. *Linking sequences between sections C'1 and K'*

An additional sequence (G)AAAUGCG has recently been located at the extreme 5'-end of section K'. However, the linking sequence between section C'1 and the above-mentioned sequence is still unknown. These non-identified sequences are estimated to encompass about 30–50 nucleotide residues and they do not contain any unique T_1 oligonucleotide (see [2]).

3.5. *Section C'2*

Section C'2 is one of the regions in 16 S RNA that is isolated with great difficulty. One subfragment recently isolated allows the reduction of the number of non-ordered oligonucleotides within this section. The previously reported pancreatic RNAase product G(G.AAG)C has been fully sequenced to give GAAGGC (spot 22) and has been located at the 5'-position of the remaining ambiguous part of the sequence contained in the brackets: (G)AAGG-(CCCCCUG,G,CG). Two alternative sequences are possible: (G)AAGGCCCCCUGGCG; (G)AAGGCGGCCCCCUG.

3.6. *Section S*

In our previous paper, there existed a discontinuity between sections C'2 and D'. We have recently characterized a sequence extension at the 3'-end of section C'2. This section, denoted S, contains the unique T_1 oligonucleotide CAAACAG (spot 93c) which was not previously found in the RNA sequence. It contains a second copy of AUUAG (spot 38) and the characteristic pancreatic product AGGAU (spot 15ii). This sequence, CAAACAGGAUUAG, is directly located at the 3'-end of section C'2.

Evidence that section S is directly adjacent to section D' is provided [5]. An adjacent T_1 oligonucleotide pair was characterized: (G)AUUAGAUACCCUG$_{OH}^{(K)}$ arising from kethoxal modification of the 16 S RNA within the 30 S subunit, the kethoxal modified G residue being denoted $G^{(K)}$. AUACCCUG (spot 28), being a unique T_1 oligonucleotide located at the 5'-end of section D', indicates that this section must be continuous with section S.

3.7. *Sections D'–O*

Very little was known about the order of oligonucleotides in section O'. However, fragments arising from this region have recently been obtained from T_1 ribonuclease digestion of salt-washed 30 S subunits, providing further data for the determination of the sequence of this section. An additional pancreatic product, GGU has been characterized (three (py)GGU copies are therefore present in the whole section instead of two). The position of (G)AUm^2G or the corresponding non methylated oligonucleotide (G)AUG has been revised and an additional CG has also been located between CAACG and AAG. The up dated sequence is shown in fig.1. Unfortunately, several ambiguities still remain in the order of the T_1 oligonucleotides contained within the brackets, the corresponding pancreatic RNase products being GGGGGC, G(A.G)GC, AAGU, GGU, GC.

3.8. *Section E'*

Partial modifications have been performed in section E'. Firstly, we have shown that (G)UCG rather than (G)CUG is directly linked to the 3'-end of section D, the corresponding pancreatic linking product being AGGU. In addition a new T_1 oligonucleotide has been shown to occur: CCUG (spot 86). Several new subfragments have allowed us to deduce a partial sequence (py)AGGUCG(CUG,UG)(CUG,G,CAUG)UCAGCCUG. . . for the 3'-part of this section.

An adjacent T_1 oligonucleotide pair (G)UCGUCAGp$^{(K)}$ located within section E' was also reported [5]. This is not in agreement with our newly characterized sequence in section E' in which UCAG can only be preceeded by CUG or CAUG, not by UCG. These authors have also characterized the adjacent T_1 oligonucleotide pair (G)UGAAAUG$_{OH}^{(K)}$ which, they

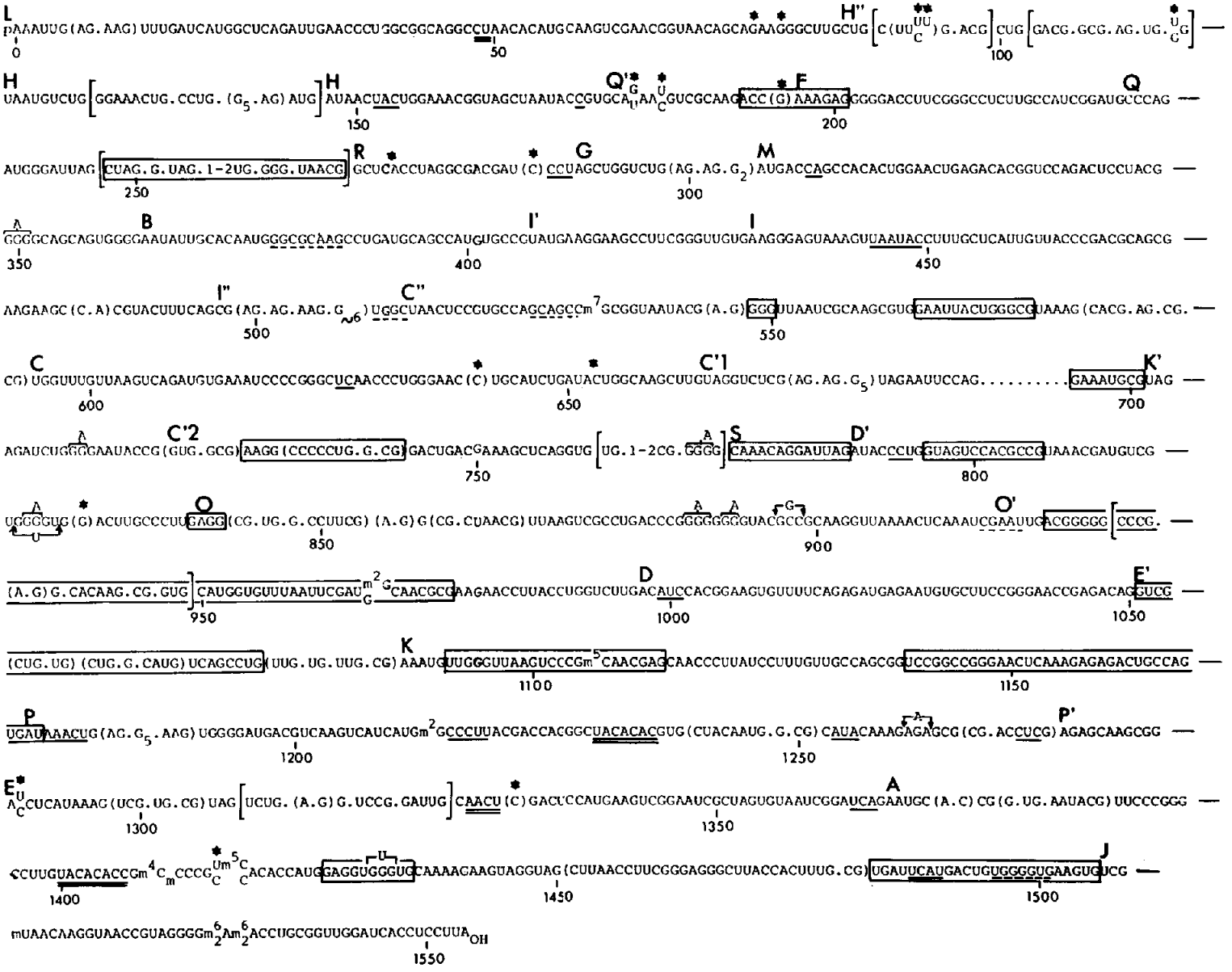

Fig.1. Primary structure of the 16 S RNA from *E. coli*. The regions in which additional information has been included are indicated by boxes (see text). The sequences of several T_1 oligonucleotides have been modified or completed: the modified sequences are underlined with a single line, the completed sequences with a double line. Stars denote cistronic heterogeneities in the sequence. Regions were precise overlap has not yet been obtained are indicated by a broken line.

propose, extends the sequence from the 5′-end of section K into section E′ [5]. The existence, however, of a second copy of (G)AAAUG at the 5′-part of section K′ makes it impossible to state precisely whether this T_1 oligonucleotide pair arises from region E′–K or region C′1–K′.

3.9. *Section K*

In our previous paper, we reported a partial or tentative sequence for section K. These ambiguities were mainly due to the difficulties encountered in obtaining fragments encompassing this section or subfragments thereof. Only recently we have succeeded in isolating a number of fragments derived from section K and we can now report the total sequence of this section (see fig.1). In the 5′-part, oligonucleotide (G)UCCCG is found to occur without (G)CCUG as previously reported. The relative positioning of (G)UUG and (G)UCCCG has been reversed to give the sequence (G)UUGGGUUAAGUCCCG. This part of sequence is in agreement with the adjacent T_1 oligo-

(K)

nucleotide pair (G)UUAAGUCCCG$_{OH}$ reported [6].

Recent work of Ross and Brimacombe (personal communication) agrees well with our proposed sequence.

3.10. *Joining K–P*

No evidence was available to show that section P was directly adjacent to section K [2] and the presence of linking sequences between these two sections could not be excluded. We have now isolated an overlapping fragment containing section P linked to the 3′-part of section K which shows unambiguously that section P is directly adjacent to the 3′-part of section K, the linking pancreatic product being GAU contained in the sequence (G)UGAUAAACUG.

The T_1 oligonucleotide pair identified as

(K)

(G)GAUAAACUG$_{OH}$ [5], corresponds to the pair 1a/32 in the designation [5]. According to our results, the T_1 product preceding AUAAACUG is UG and not G, the corresponding pancreatic product being GAU and not XGGAU. After examination of the results [5], we believe that spot 17, corresponding to UGp rather than spot 1a (Gp), could be the oligonucleotide adjacent to spot 32.

3.11. *Section A*

The isolation of a new T_1 partial fragment from section A, namely m^4C_mCCCGUm^5 CACACCAUGGAGGU(GGG.U)UGCAAAAG, led us to check the sequence of this region. It appears that the order of the T_1 oligonucleotides: AUUCAUG (spot 13), ACUG (spot 78) and two UG copies and their corresponding pancreatic oligonucleotides: GAU, GAC, GGGGU, and two GU copies have been reversed with regard to the central sequence CAAAAGAAGUAGGUAG(CUUAACCUUCGGGA-GGGCUUACCACUUUG,CG). The revised sequence is shown fig.1. An additional sequence (G)AAGUG has been found to occur at the 3′-end of section A, corresponding to the pancreatic product GAAGU. The 3′-part (adjacent to section J) is only tentative because no formal overlap has been isolated for this region.

Although this reordering agrees with new results of Ross and Brimacombe (personal communication) several discrepancies still remain in the sequences located at the 3′-side of GGGAGGGC.

3.12. *Oligonucleotide sequence revision*

A number of minor modifications have been reported [7,8] concerning the internal sequences of several T_1 oligonucleotides. The use of U_2 ribonuclease and subsequent appropriate analysis has enabled us to confirm the sequences reported [7,8]. Some of these modifications have already been taken into account [2]. Such modified or completed sequences are indicated in fig.1.

4. Discussion

The up-dated sequence of the 16 S RNA encompasses 1554 nucleotides, 1350 of which are contained within regions of complete sequence. One gap remains within the sequence between sections C′1 and K′ in which all the unidentified sequences (about 30–50 nucleotides) must be located. Figure 1 indicates the regions in which additional or modified data have been introduced. Several ambiguities are still present in different regions throughout the molecule. The extreme difficulty encountered during the determination of the primary sequence of several areas of the 16 S RNA reflects the limits imposed by the finger-

printing method as well as the consequences of the structural organization of the 16 S RNA: several regions are very exposed and susceptible to partial ribonuclease digestion in the free 16 S RNA as well as within the 30 S subunit [9]. Therefore, these regions are very difficult to isolate. This is particularly the case for the unknown region located between sections C'1 and K' which has never been isolated.

The secondary structure and several features of tertiary organization of the molecule will be discussed elsewhere [9]. We now believe that the most direct method with which to complete the primary sequence is by the use of the techniques recently developed using DNA [10] and using RNA [11]. This work is now in progress.

Acknowledgements

We are endebted to Drs Ross and Brimacombe for making available their last results concerning the 16 S RNAs sequence. This work was supported financially by the Délégation à la Recherche Scientifique et Technique and the Commissariat à l'Energie Atomique.

References

[1] Ehresmann, C., Stiegler, P., Mackie, G. A., Zimmermann, R. A., Ebel, J. P. and Fellner, P. (1975) Nucleic Acids Res. 2, 265–278.
[2] Ehresmann, C., Stiegler, P., Fellner, P. and Ebel, J. P. (1975) Biochimie 57, 711–748.
[3] Sanger, F., Brownlee, G. G. and Barrel, B. G. (1965) J. Mol. Biol. 13, 373–398.
[4] Ehresmann, C., Stiegler, P. and Ebel, J. P. (1974) FEBS Lett. 49, 47–48.
[5] Chapman, N. M. and Noller, H. F. (1977) J. Mol. Biol. 109, 134–149.
[6] Noller, H. F. (1974) Biochemistry 13, 4694–4703.
[7] Uchida, T., Bonen, L., Shaup, H. W., Lewis, B. J., Zablen, L. and Woese, C. (1974) J. Mol. Evol. 3, 63–77.
[8] Magrum, L., Zablen, L., Stahl, D. and Woese, C. (1975) Nature 257, 423–426.
[9] Ehresmann, C., Stiegler, P., Carbon, P. and Ebel, J. P. manuscript in preparation.
[10] Maxam, A. M. and Gilbert, W. (1977) Proc. Natl. Acad. Sci. USA 74, 560–564.
[11] Donis-Keller, H., Maxam, A. M. and Gilbert, W. (1977) Nucleic Acids Res. 4, 2527–2538.