Volume 249, number 2, 279–284 FEB 07206 June 1989

# Structure of the 5′-external transcribed spacer of the human ribosomal RNA gene

Marie-Hélène Renalier, Sylvie Mazan, Nicole Joseph, Bernard Michot and Jean-Pierre Bachellerie

*Centre de Recherche de Biochimie et Génétique Cellulaires du CNRS, Université Paul-Sabatier, 118, route de Narbonne, 31062 Toulouse Cédex, France*

Received 29 March 1989; revised version received 13 April 1989

We report the complete nucleotide sequence of the 3627 bp long 5′-external transcribed spacer (ETS) of a human ribosomal RNA gene. This sequence exhibits only very limited homologies with its mouse counterpart, the only other mammalian specimen analyzed so far. It has very peculiar compositional characteristics, with a highly biased base content (very rich in G+C, very poor in A) and also some very strong dinucleotide preferences. Interestingly, these specific features are shared by the mouse sequence, despite the extensive sequence divergence, and also apply to the other transcribed spacers of mammals indicating that a common and strong structural constraint is exerted on all these regions of the ribosomal gene. An outstanding secondary structure can be formed within the human ETS RNA, which could have a significant role in preribosome assembly.

rRNA gene; External transcribed spacer; Nucleotide sequence; Secondary structure; Compositional constraint; Dinucleotide frequency

## 1. INTRODUCTION

In eucaryotes, mature rRNAs are processed in a stepwise fashion from a large primary transcript in which spacer sequences alternate with 18 S, 5.8 S and 28 S rRNAs [1]. Spacer sequences constitute a large part of the transcription unit in higher eucaryotes, particularly in mammals for which they amount to about half of the 13–14 kb gene length [2]. Their potential roles in control of ribosome biogenesis remain speculative so far. As for rRNA processing reactions, their molecular mechanisms are not known and the structures in pre-rRNA required for the recognition of the proper cleavage sites have not been identified either, except for the primary processing event in mammals which occurs within the 5′-portion of the 5′-external transcribed spacer [3–5]. Sequences of ribosomal transcribed spacers undergo a fast divergence during evolution [6–11], in contrast to mature rRNAs. As for the 5′-external transcribed spacer, which is roughly 4 kb long in mammals, the only vertebrate sequences reported so far corresponded to *Xenopus* [9,10] and mouse [11], which appear extensively divergent from each other. In the present study, we present the complete sequence, and a model of the secondary structure, for the human 5′-external transcribed spacer (5′-ETS). Although the two mammalian sequences now available for comparison exhibit only a very low homology, they appear to be subject to common and specific compositional constraints.

*Correspondence address:* J.-P. Bachellerie, Centre de Recherche de Biochimie et Génétique Cellulaires du CNRS, Université Paul-Sabatier, 118, route de Narbonne, 31062 Toulouse Cédex, France

The nucleotide sequence presented here has been submitted to the EMBL/GenBank database under accession no. X14345

## 2. MATERIALS AND METHODS

Several complete ribosomal genes have been isolated [12] from a human cosmid library, constructed from leucocyte DNA using pCV 108 as cosmid vector [13]. The 5.8 kb *Eco*RI fragment of human rDNA (which contains a 0.5 kb sequence upstream from the transcription start site, the entire 5′-ETS and most of the 18 S rRNA coding region) was isolated from the cosmid insert of the 11.9 clone [12] and subcloned in pUC 8.

*Published by Elsevier Science Publishers B.V. (Biomedical Division)*
00145793/89/$3.50 

Different fragments were subcloned again in pUC 8 and M13 [14] vectors, and sequenced by the dideoxynucleotide chain termination method [15]. No ambiguity remained over the entire sequence which was determined on both strands over most of its length. The sequences of the (+1, +700) region for this clone and for other different human genes had been reported previously [4,12,16].

RNA secondary structure predictions [17] were performed on the entire sequence by serial analysis of portions measuring up to 2.2 kb. Sequence analyses were carried out with the software package [18] of the Genetics Computer Group, University of Wisconsin.

## 3. RESULTS AND DISCUSSION

The 3627 nucleotide long human 5'-ETS sequence is very rich in G + C (79%) and particular-

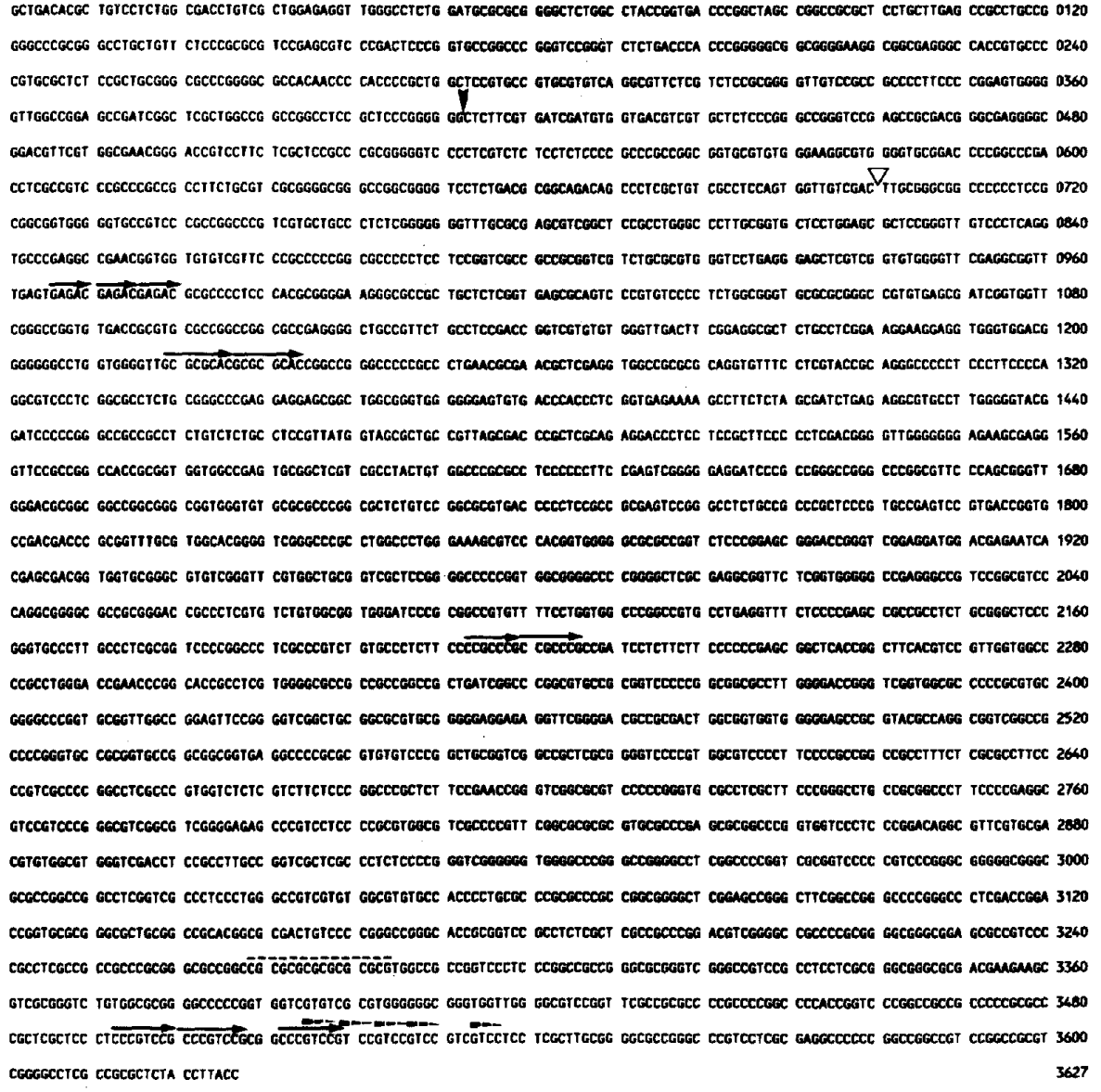

Fig.1. Sequence of the human external transcribed spacer. The sequence extends from the transcription start site (position +1) to the nucleotide (position +3627) which immediately precedes the mature 18 S rRNA coding sequence. The site of internal processing of the 5'-ETS is denoted by a solid arrowhead and the 3'-boundary of the previously sequenced leader segment by an open triangle. The most outstanding sets of tandemly repeated oligonucleotides are overlined by arrows, and a run of alternating CG by dots.

ly poor in A (5.65%), these overall values reflecting a rather uniform base content along its entire length. The sequence does not exhibit any extended pattern of internal repetition. As for tandem repetitions of short oligonucleotides (overlined in fig.1), their number does not significantly exceed that which would be expected on a random basis forany sequence of the same base composition, suggesting that DNA-strand slippages during replication [19] are not frequent events relative to the rate of nucleotide substitution of this sequence.

According to the Fickett's criteria [20], the sequence appears devoid of any likely open reading frame. Comparison with its mouse counterpart [11], which is only slightly longer, reveals a very limited extent of sequence conservation (fig.2), with only seven tracts displaying a substantial homology (fig.2b). Tract 1, identified in a previous analysis of the 5′-end of the human ETS [4], extends over about 200 bp immediately downstream from the site of internal ETS processing. Tracts 2–7, which occur in the same linear order in both mammalian sequences, are shorter but the homology appears significant. However this conservation does not extend to distant vertebrates, since no residual homology could be detected over these tracts with the amphibian 5′-ETS sequences [9,10]. No significant match was found between the human 5′-ETS and any of the nucleotide sequences in GenBank.

Remarkably, the human and mouse 5′-ETS sequences, although extensively divergent, are closely related in their highly biased base content. While the average GC content in mouse and human genomes is around 42% [21], the two mammalian 5′-ETS are dramatically enriched in GC (fig.3a,b). In each case, the content in G is nearly identical to the content in C. Moreover both sequences are also extremely poor in A, displaying the same very strong imbalance in U and A contents (the U/A ratio averages 2.70 in both species). It is intriguing that all these unusual compositional characteristics also apply to the internal transcribed spacer regions of ribosomal genes in these two mammals (fig.3c,d). In contrast, they are definitely not shared by the mature rRNA coding regions (fig.3e).

Another kind of strong structural constraint operating on the human 5′-ETS is revealed by the examination of dinucleotide frequencies. As shown

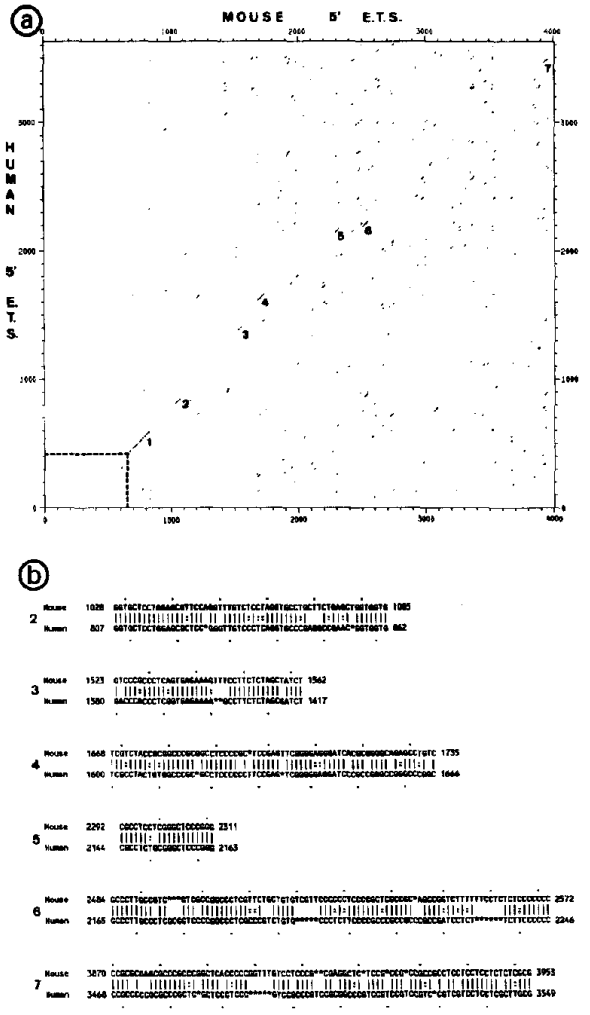


Fig.2. Comparison of the human and mouse 5′-ETS sequences. (a) Homology matrix: each dot denotes the presence in both sequences of a 32 nucleotide long segment displaying more than 65% homology. The location of the site of internal 5′-ETS processing is indicated by a broken line. Positions are numbered from the transcription start site. (b) Details of sequence alignment for the longest homologous tracts (strings of numbered contiguous diagonal dots in a). The previously identified [4] tract 1 is not shown. Identities are denoted by bars and transition mismatches by colons.

in fig.4, the values observed for some doublets strongly depart from what would be expected for a randomized sequence of the same base composition. This is particularly dramatic for the direct environment of As: as a 5′-neighbor, A and G are highly preferred, whereas, as a 3′-neighbor, T is markedly unfavorable (fig.4a). Strikingly, the

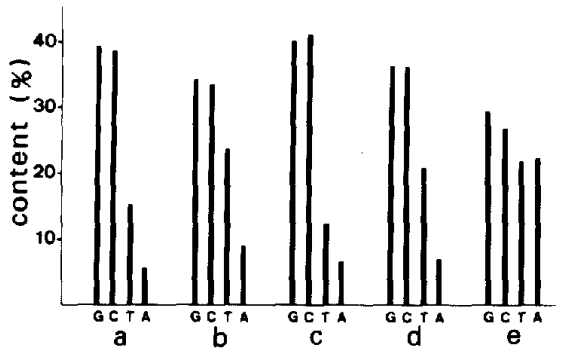


Fig.3. Base composition of the mammalian ribosomal transcribed spacers. (a) Human 5′-ETS; (b) mouse 5′-ETS [11]; (c) human ITS 1 + ITS 2 (Sylvester, J.E., personal communication); (d) mouse ITS 1 + ITS 2 [8]; (e) human or mouse [26] 18 S rRNA.

same peculiarities are also observed for the external and internal transcribed spacers of the mouse ribosomal gene (fig.4b,c). However, they clearly do not extend to the mature 18 S rRNA coding region (fig.4d) which displays a much more even distribution of dinucleotide frequencies. This close similarity, in terms of dinucleotide preferences, among mammalian ribosomal transcribed spacers is not restricted solely to the above-mentioned cases: in fact, the five most preferred doublets remain the same, i.e. AA, GA, TC, GT and AG (in the same order or nearly so), in each sequence. Dinucleotide frequencies in DNA or RNA sequences do not fluctuate randomly and recurring patterns of preference have been observed in some phylogenetic groups which might be indicative of steric constraints on the conformation and packaging of the double helix [22]. In this regard, the strong preferences shared by the mammalian ribosomal transcribed spacers might reflect a peculiar chromatin organization of these portions of the gene as compared to the mature rRNA coding regions, in line with recent data on the structure of actively transcribed ribosomal chromatin in *Xenopus* [23]. Accordingly, these intragenic spacers could well have a role as DNA structures, possibly involved in a control of the transcription elongation process. Nevertheless their potential importance as RNA structures should not be overlooked, since their peculiar compositional features are correlated with the likely appearance of exceptionally stable secondary structures in mammals. A giant hairpin loop had been detected in the external spacer region of mammalian pre-rRNAs by electron microscopic secondary structure mapping [24,25]. The organization of this hairpin varies among mammals, with changes in the number and lengths of branches, although the overall size is maintained. The human 5′-ETS sequence was systematically searched for the most favorable secondary structure and two highly stable domains were reproducibly found (fig.5). The larger one (fig.5b) unambiguously cor-

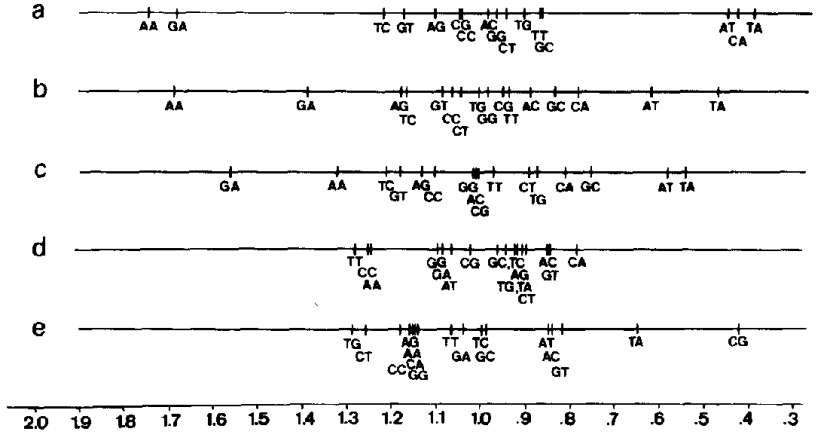


Fig.4. Frequencies of occurrence of the dinucleotides. For each dinucleotide, values were obtained for the corresponding RNA-like strand by dividing the observed number by the number expected on a random basis from the overall base content of each sequence: human 5′-ETS (a); mouse 5′-ETS (b); mouse internal transcribed spacers 1 + 2 (c); mouse 18 S rRNA (d); averaged frequencies for a compilation [22] of vertebrate sequences (e).

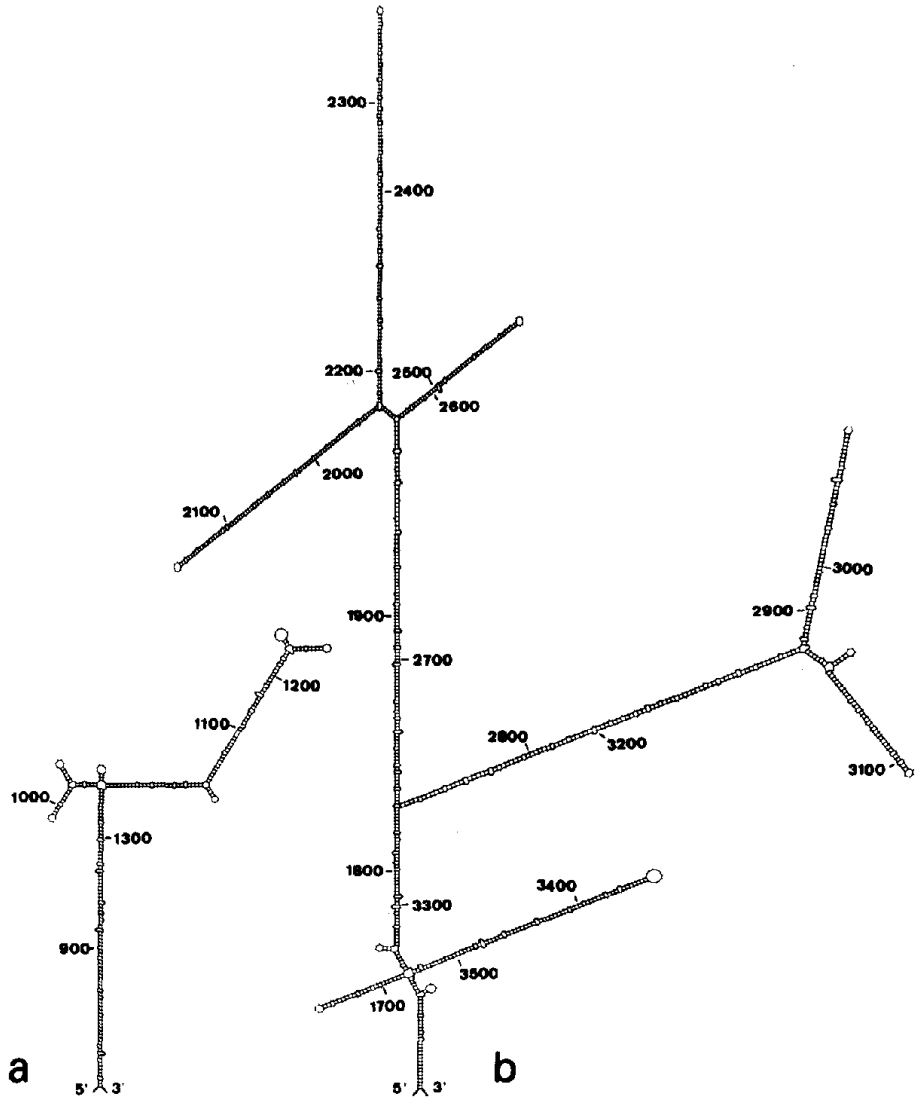


Fig.5. Secondary structure of the human 5'-ETS. Parts of the thermodynamically favored folding pattern are represented, for domains 851–1391 (a) and 1657–3563 (b) of the human sequence.

responds to a characteristic structure observed in primates by electron microscopy (plate IIId in [24] and fig.1H in [25]) not only in size and location but also in details of its branched organization. It encompasses a major portion of the 5'-ETS sequence (i.e. 1907 nucleotides) and maps only 60 nucleotides upstream from the 5'-end of the 18 S rRNA coding region in human pre-RNA. The refolding of this giant structure during the elongation of nascent pre-RNA could mediate major conformational switches required in the early steps of preribosome assembly.

*Acknowledgements:* The support of Professor J.P. Zalta was appreciated. This work was aided by grants from INSERM (CRE no. 851.001) and from ARC (no. 6158).

## REFERENCES

[1] Perry, R.P. (1976) Annu. Rev. Biochem. 45, 605–629.

[2] Crouch, R.J. and Bachellerie, J.P. (1986) Ribosomal RNA processing sites. In: DNA systematics vol. I (Dutta, S.K. ed.) CRC Press, Boca Raton, FL, pp. 47–80.
[3] Gurney, T., jr (1985) Nucleic Acids Res. 13, 4905–4919.
[4] Kass, S., Craig, N. and Sollner-Webb, B. (1987) Mol. Cell. Biol. 7, 2891–2898.
[5] Craig, N., Kass, S. and Sollner-Webb, B. (1987) Proc. Natl. Acad. Sci. USA 84, 629–633.
[6] Hall, L.M.C. and Maden, B.E.H. (1980) Nucleic Acids Res. 8, 5993–6005.
[7] Furlong, J.C. and Maden, B.E.H. (1983) EMBO J. 2, 443–448.
[8] Michot, B., Bachellerie, J.P. and Raynal, F. (1983) Nucleic Acids Res. 11, 3375–3391.
[9] Maden, B.E.H., Moss, M. and Salim, M. (1982) Nucleic Acids Res. 10, 2387–2398.
[10] Furlong, J.C., Forbes, J., Robertson, M. and Maden, B.E.H. (1983) Nucleic Acids Res. 11, 8183–8196.
[11] Bourbon, H., Michot, B., Hassouna, N., Feliu, J. and Bachellerie, J.P. (1988) DNA 7, 181–191.
[12] Rénalier, M.H., Joseph, N. and Bachellerie, J.P. (1989) FEBS Lett., in press.
[13] Lau, Y.F. and Kan, Y.W. (1983) Proc. Natl. Acad. Sci. USA 80, 5225–5229.
[14] Messing, J., Crea, R. and Seeburg, P.H. (1981) Nucleic Acids Res. 9, 309–321.
[15] Sanger, F., Nicklen, S. and Coulson, A. (1977) Proc. Natl. Acad. Sci. USA 74, 5463–5467.
[16] Financsek, I., Mizumoto, K., Mishima, Y. and Muramatsu, M. (1982) Proc. Natl. Acad. Sci. USA 79, 3092–3096.
[17] Zuker, M. and Stiegler, P. (1981) Nucleic Acids Res. 9, 133–148.
[18] Devereux, J., Haeberli, P. and Smithies, O. (1984) Nucleic Acids Res. 12, 387–395.
[19] Levinson, G. and Gutman, G.A. (1987) Mol. Biol. Evol. 4, 203–221.
[20] Fickett, J.W. (1982) Nucleic Acids Res. 10, 5303–5318.
[21] Nussinov, R. (1987) DNA 6, 13–22.
[22] Nussinov, R. (1984) Nucleic Acids Res. 12, 1749–1763.
[23] Culotta, V. and Sollner-Webb, B. (1988) Cell 52, 585–597.
[24] Wellauer, P.K., Dawid, I.B., Kelley, D.E. and Perry, R.P. (1974) J. Mol. Biol. 89, 397–407.
[25] Schibler, U., Wyler, T. and Hagenbüchle, O. (1975) J. Mol. Biol. 94, 503–517.
[26] Raynal, F., Michot, B. and Bachellerie, J.P. (1984) FEBS Lett. 167, 263–268.