Hypothesis

5 S RIBOSOMAL RNA GENES AND THE *Alu*I FAMILY: EVOLUTIONARY AND FUNCTIONAL SIGNIFICANCE OF A REGION OF STRONG HOMOLOGY

W. Ford DOOLITTLE
Department of Biochemistry, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada

Received 5 January 1981

The transcription of a number of low-molecular-weight cellular and viral RNAs is catalyzed by RNA polymerase III [1]. In 3 instances, intragenic sequences have been defined as essential for transcription initiation: nucleotides between (mature RNA) positions 50–83 in a *Xenopus* 5 S gene; between positions 15–56 in a *Xenopus* $tRNA^{tyr}$ gene; and between positions 9–72 in the adenovirus VAI RNA gene [1–8].

Comparisons of intragenic sequences implicated in such studies should lead to an identification of a common recognition site(s), but have met with limited success. Bogenhagen et al. [3] noted that the *Xenopus* 5 S sequence AGCAGGGU (positions 55–62) has partial homologs in VAI RNA, *Bombyx mori* $tRNA^{ala}$ and yeast $tRNA^{tyr}$. However, Fowlkes and Shenk [8] defined two different regions of partial homology in VAI(A) and VAII adenovirus RNAs, several tRNAs and mouse 4.5 S RNA. These regions are not present in *Xenopus* 5 S, and attempts by Koski et al. [6] to define sequences common to a variety of tRNAs and 5 S rRNAs revealed only 9 fully conserved and 5 partially conserved residues scattered over 2 regions which together comprise some 39 positions. The identification of a 38 000 M_r protein which stimulates the transcription of cloned 5 S rRNA genes but not cloned tRNA genes [5,9] supports the notion that there are at least 2 classes of intragenic RNA polymerase III initiation sites, recognized by different (dissociable?) protein factors [8].

Members of the *Alu*I family of mammalian repetitive DNAs [10,11] contain an RNA polymerase III recognition site(s) whose position(s) remains unknown. In an effort to find this site(s), I have compared the sequences of 4 independently-cloned members of the human *Alu*I family [10,12,13] with the sequence of human (KB) 5 S rRNA in a region (residues 45–87) expected to encompass the RNA polymerase III recognition site. The best alignment is shown in fig.1, where the *Alu*I site of each member (which lies near the center of each 300 basepair element) is positioned beneath residues 74–76 of KB 5 S rRNA. Two 4-residue and 2 single residue 'additions' were assumed (at identical positions) for all *Alu*I family members, and these positions (indicated by hyphens in the 5 S sequence) were ignored in all subsequent calculations. Residues present in KB 5 S rRNA which are common to 6 other eukaryotic 5 S rRNAs chosen on the basis of phylogenetic and sequence divergence (those of *Xenopus laevis* somatic cells [14], *Drosophila melanogaster* [15], *Torulopsis utilis* [14], *Triticum vulgare* [16], *Tetrahymena thermophila* [17] and *Crithidia fasciculata* [18]) are indicated by capital letters, as are residues common to the 4 *Alu*I clones. From fig.1, I draw the following conclusions and inferences.

1. Homology between KB 5 S rRNA and the *Alu*I family members begins abruptly at 5 S residue 50 and ends at residue 83 or 85; no substantial sequence similarities are detectable outside this region. Within it, however, homologies are surprisingly strong. *Alu*I clones A,B,C and D, respectively, show 29, 25, 24 and 25 residues (indicated by underlining) in common with the 36 residue region of KB 5 S rRNA, yielding sequence homologies from 66.7–80.6% (identical over total), with an average value of 71.5%. Nineteen of the 36 residues are common to all 5 sequences shown in fig.1

2. Pairwise comparisons between the 4 *Alu*I family members within the 36-residue region show homologies ranging from 66.7–88.9% (av. 79.2%). These values are not dramatically higher than those ob-

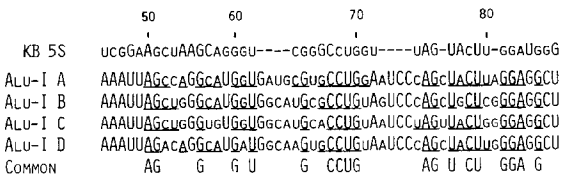

Fig.1. The sequence of human KB 5 S rRNA (residues 45–87) aligned against the sequence of 4 independently cloned members of the human *Alu*I family. *Alu*IA is 'clone 8' [10], *Alu*IB is in the 5′-flanking region of the human ε-globin gene [12], *Alu*IC was found as an insert in a mutant of SV40 [13], and *Alu*ID is in the 5′-flanking region of the human G-γ-globin gene (data of S. M. Weissman, cited in [13]).

tained between *Alu*I family members and KB 5 S rRNA, it is as if evolutionary pressures to keep *Alu*I family members homologous to 5 S rRNA in this region are comparable to pressures to keep them homologous to each other. The ~300 000 members of the human *Alu*I family are each ~300 basepairs long and on average show an overall sequence homogeneity of 76–88% [10]. One must conclude either that selective pressures operating on the remaining 260–270 residues are also of comparable strength, or that some form of correction and remultiplication mechanism (e.g., transposition) serves to keep family members from diverging greatly in sequence [19–21].

3. If residues U_{72} and U_{73} are excluded (because they may have been displaced in some insertion event), there are 21 5 S positions between residues 50–85 which are not identical in KB 5 S rRNA and the 6 other diverse 5 S rRNAs listed above. Each of the 4 *Alu*I clones can be scored for its homology in each of these 21 positions with each of the other 5 S rRNAs. On average, the clones share 15.8, 15.8, 10.0, 5.8, 7.3, 12.3 and 6.0 of the 21 positions with the 5 S rRNAs of human (KB), *X. laevis, D. melanogaster, Torulopsis utilis, Tetrahymena thermophila, Triticum vulgare* and *C. fasciculata,* respectively. In other words, in positions which are not common to all 7 5 S rRNAs, *Alu*I family members show strongest homology (75.2%) to vertebrate 5 S rRNAs, lower values with insect and plant 5 S, and nearly random values (30.3%) with protozoan and fungal 5 S rRNAs. KB 5 S rRNA itself shows roughly the same pattern, sharing out of the 21 positions, 20, 15, 8, 9, 14 and 9 nucleotides, respectively, with the 5 S rRNAs of *X. laevis, D. melanogaster, Torulopsis utilis, Tetrahymena thermophila, Triticum vulgare* and *C. fasciculata.* These observations suggest that either:

(i) Constraints on the region common to 5 S rRNA and the *Alu*I family members are rather loose, and the closer homology of the *Alu*I sequence region identified here to that of vertebrate 5 S regions reflects its derivation from vertebrate 5 S rRNA genes; or

(ii) Constraints on these regions vary among eukaryotes, and the *Alu*I sequence has been selected to conform to vertebrate-specific constraints.

4. The two 'insertions' (between 5 S residues 62 and 63 and 72 and 73) assumed to have occurred in the *Alu*I family effectively divide the region of 5 S *Alu*I homology into 3 regions. These correspond almost exactly to the 3 regions of *Xenopus* XBS1 5 S rDNA found [9] to be protected by the 38 000 M_r protein which specifically stimulates synthesis of 5 S rRNA.

On the basis of these observations, it seems reasonable to suggest that:

(i) The regions of homology between KB 5 S rRNA and members of the *Alu*I family identified here represent recognition site for RNA polymerase III;

(ii) Both are recognized by the same DNA-binding protein;

(iii) Either the *Alu*I sequence derived directly from a segment of vertebrate 5 S rRNA genes, or that intragenic 5 S rDNA recognition sites (and the corresponding proteins) vary within the eukaryotes.

Each of these suggestions is testable. What remain unresolved are the actual site(s) or RNA polymerase III transcription initiation in *Alu*I DNAs, the mechanism by which complementary transcripts of both strands of the DNA are produced [10,11], and the function of such transcripts either in the maintenance of the *Alu*I family or in other, selectively advantageous, cellular processes.

References

[1] Ford, P. J. (1980) Nature 287, 109–110.
[2] Sakonju, S., Bogenhagen, D. F. and Brown, D. D. (1980) Cell 19, 13–25.
[3] Bogenhagen, D. F., Sakonju, S. and Brown, D. D. (1980) Cell 19, 27–35.
[4] Pelham, H. R. B. and Brown, D. D. (1980) Proc. Natl. Acad. Sci. USA 77, 4170–4174.
[5] Telford, J. L., Kressmann, A., Koski, R. A., Grosschedl, R., Müller, F., Clarkson, S. G. and Birnstiel, M. L. (1979) Proc. Natl. Acad. Sci. USA 76, 2590–2594.
[6] Koski, R. A., Clarkson, S. G., Kurjan, J., Hall, B. D. and Smith, M. (1980) Cell 22, 415–425.
[7] Akusjärvi, G., Mathews, M. B., Andersson, P., Vennström, B. and Pettersson, U. (1980) Proc. Natl. Acad. Sci. USA 77, 2424–2428.
[8] Fowlkes, D. M. and Shenk, T. (1980) Cell 22, 405–413.
[9] Engelke, D. R., Ng, S.-Y., Shastry, B. S. and Roeder, R. G. (1980) Cell 19, 717–728.
[10] Rubin, C. M., Houck, C. M., Deininger, P. L., Friedmann, T. and Schmid, C. W. (1980) Nature 284, 372–374.
[11] Jelinek, W. R., Toomey, T. P., Leinwand, L., Duncan, C. H., Biro, P. A., Choudary, P. V., Weissman, S. M., Rubin, C. M., Houck, C. M., Deininger, P. L. and Schmid, C. W. (1980) Proc. Natl. Acad. Sci. USA 77, 1398–1402.
[12] Baralle, F. E., Shoulders, C. C., Goodbourn, S., Jeffrys, A. and Proudfoot, N. J. (1980) Nucleic Acids Res. 8, 4393–4404.
[13] Dhruva, B. R., Shenk, T. and Subramanian, K. N. (1980) Proc. Natl. Acad. Sci. USA 77, 4514–4518.
[14] Erdmann, V. A. (1980) Nucleic Acids Res. 8, r31–r47.
[15] Tschudi, C. and Pirrotta, V. (1980) Nucleic Acids Res. 8, 441–451.
[16] MacKay, R. M., Spencer, D. F., Doolittle, W. F. and Gray, M. W. (1981) Eur. J. Biochem. in press.
[17] Luehrsen, K. R., Fox, G. E. and Woese, C. R. (1981) Curr. Microbiol. in press.
[18] Mackay, R. M., Gray, M. W. and Doolittle, W. F. (1980) Nucleic Acids Res. 8, 4911–4917.
[19] Klein, W. H., Thomas, T. L., Lai, C., Scheller, R. H., Britten, R. J. and Davidson, E. H. (1978) Cell. 14, 889–900.
[20] Doolittle, W. F. and Sapienza, C. (1980) Nature 284, 601–603.
[21] Orgel, L. E. and Crick, F. H. C. (1980) Nature 284, 604–607.