Volume 13 Number 14 1985 Nucleic Acids Research

An unusually long poly(purine)-poly(pyrimidine) sequence is located upstream from the human thyroglobulin gene

D.Christophe, B.Cabrer, A.Bacolla, H.Targovnik, V.Pohl+ and G.Vassart*

Institut de Recherche Interdisciplinaire (IRIBHN), *Service de Chimie and +Service d'Histologie, Faculté de Médecine, Université Libre de Bruxelles, Campus Erasme, route de Lennik 808, B-1070 Bruxelles, Belgium

Received 23 April 1985; Revised and Accepted 25 June 1985

ABSTRACT

A region of human genomic DNA encompassing the 5' end of the thyroglobulin gene has been sequenced and the position of the transcriptional start site has been determined. The 5' non-translated portion of the mRNA displays a quasi-palindromic sequence which could allow this region to adopt a hairpin structure. The first exon of the gene encodes a 19 amino-acids signal peptide and the 3 first amino acids of the mature protein. Apart from the canonical TATA-Box and from a CAAT-Box homology, the promoter region contains a 209 bp-long poly(purine)-poly (pyrimidine) sequence located between positions-512 and -304 relative to the transcription start. When contained in a supercoiled plasmid, this sequence exhibits sensitivity to S1 nuclease at two distinct positions. A precise mapping of the borders of the sensitive regions was achieved by extending primers from both ends of the sequence after digestion by the enzyme. The resulting data can be explained by a model involving the formation of a triple helix structure.

INTRODUCTION

With an estimated size of about 200 kb, the thyroglobulin (Tg) gene appears to be one of the largest genes known to date (1). This giant transcription unit encodes a 660 kD protein made of two identical subunits (2). So far, the only known role for Tg is that of precursor of the thyroid hormones, thyroxine and 3,5,3'-triiodothyronine, which are formed following iodination and oxidative coupling of only a few ($\simeq$10) out of the 144 tyrosine residues of the protein (3). This process involves well defined positions in Tg sequence, called hormonogenic sites, four of which have been characterized recently (4,5).

Tg synthesis in the thyroid is under control of the pituitary hormone thyrotropin (TSH), the action of which is mediated by cyclic AMP (cAMP) (6). By direct measurement of the transcription rate of the rat Tg gene, Van Heuverswyn et al. (7) recently demonstrated that this control is mainly transcriptional. Their results also indicated that the rat Tg gene is already maximally expressed in the presence of physiological concentrations of TSH.

The continuous presence of TSH is required to maintain the gene in an activated state, as decreasing TSH levels dramatically decreases its transcription. Since the TSH effect on Tg gene expression is mimicked by Forskolin, an agent known to increase cAMP level in thyroid cells (8), this strongly supports the concept that the effective modulator of Tg gene transcription is cAMP (9).

In contrast to the widely studied mechanism of the control of gene expression by the steroid hormones (10, 11), the way by which cAMP regulates transcription in eukaryotes is still unknown. This is mainly due to the fact that only a few genes have been identified the transcription of which is controlled by a cAMP-dependent mechanism (12, 13, 14). Therefore, regulation of Tg gene expression in thyroid cells provides a good model to study at the molecular level the general phenomenon of transcription control by cAMP in eukaryotes. As a first step towards this goal, we report here the cloning and physical characterization of the human Tg gene promoter region.

MATERIALS AND METHODS

Enzymes and nucleotides : Restriction enzymes were from Boehringer-Mannheim, BRL and Amersham ; bacterial alkaline phosphatase, T4 polynucleotide kinase and T_4 DNA ligase were from BRL; S1 nuclease and the large fragment of DNA-polymerase I were from Boehringer-Mannheim; AMV reverse transcriptase was from P.H. Stehelin & cie AG; Mung Bean Nuclease; nucleotides and sequencing primers were from P.L. Biochemical; γ-^{32}P-dATP (800 Ci/mmole, 10 mCi/ml aqueous solution) and α -^{32}P-ATP (5,000 Ci/mmole, 10 mCi/ml aqueous solution) were from Amersham.

Clones and DNA sequencing : standard recombinant DNA manipulations were conducted according to (15). DNA sequences were obtained from M13 clones using the chain terminator method (16) and analyzed on a Wang VS50A computer using home-made programs.

Primer extension on RNA templates. Total RNAs were extracted using the LiCl-urea technique (17). The synthetic oligodeoxynucleotide (see Results) was labeled at its 5'end using T_4 polynucleotide kinase and γ-^{32}P-ATP according to BRL's recommendations. The labeling mixture was extracted once with phenol and twice with ether and aliquots containing approximatively 20 ng of end-labeled primer were used to redissolve 100 μg of ethanol precipitated total RNA from either human thyroid or canine sertolioma. After heating at 70°C for 15 min, the mixture was allowed to cool at room temperature for 20 min and the nucleic acids were ethanol precipitated. The pellet was

redissolved in the reverse transcription mixture (10 µl) which contained 50 mM Tris-HCl pH 8.5, 50 mM KCl, 10 mM $MgCl_2$, 40 mM dithiothreitol, all four dNTPs at 1 mM each and 38 units of reverse transcriptase. After incubation at 37° for 30 min, 5 µl of dye-containing formamide were added. The mixture was heated at 95°C for 5 min and immediately loaded on a 8 % acrylamide denaturing gel.

Digestions with S1 nuclease : the reactions were conducted at 37° for 15 min in a 50 µl volume containing 50 mM sodium acetate pH 4.5, 0.3 M NaCl, 4 mM $ZnSO_4$, 20 µg/ml t-RNA, 2 µg DNA and 800 units of enzyme. The digestions were stopped by addition of an equal volume of Tris-HCl 0.5 M pH 8, EDTA 5 mM and phenol extracted. After an ethanol precipitation step, the DNAs were submitted to restriction enzyme analysis or used as template for primer extension (see below).

Digestions with Mung Bean Nuclease : 2 µg of DNA were incubated at 37 ° for 15 min with 70 units of the enzyme in 50 mM sodium acetate adjusted to pH 4.5, 5 or 6, 0.3 M NaCl, 2 mM $ZnSO_4$ and 0.02 % Triton. Further processing was as described above.

Primer extension on DNA templates about 1 µg of S_1 nuclease-treated DNA and 1.2 ng of sequencing primer (or 4 ng of reverse sequencing primer) contained in a 8 µl volume of water in a sealed capillary were heated at 100° for 3 min. and immediately transfered in a dry-ice/ethanol bath. The mixture was allowed to thaw at room temperature and was then blowed out in a tube containing 1 µl of 100 mM Tris-HCl pH 8.5, 100 mM $MgCl_2$. After further incubation at room temperature for 20 min, the following reagents were added : 1 µl of α-^{32}P-dATP, 1 µl of a solution containing dCTP, dGTP and dTTP at 0.66 mM each, and 0.5 µl (2.5 units) of the large fragment of DNA-polymerase I. The reaction was allowed to proceed at room temperature for 15 min, then 2 µl of 0.5 mM dATP were added and the mixture was further incubated during 15 min in the same conditions. After addition of 4 µl of dye-containing formamide, the mixture was heated at 95° for 5 min and immediately loaded on a 6 % acrylamide sequencing gel.

RESULTS

Sequencing of the 5' end of the human Tg gene. The isolation and characterization of a recombinant plasmid containing a 7.9-kb insertion encompassing the beginning of the human Tg gene has been described previously (18). To study the 5' end of the gene, the internal 3.3-kb Bam HI fragment, containing only the first exon and about 2,5 kb of sequence upstream from the gene, was sub-

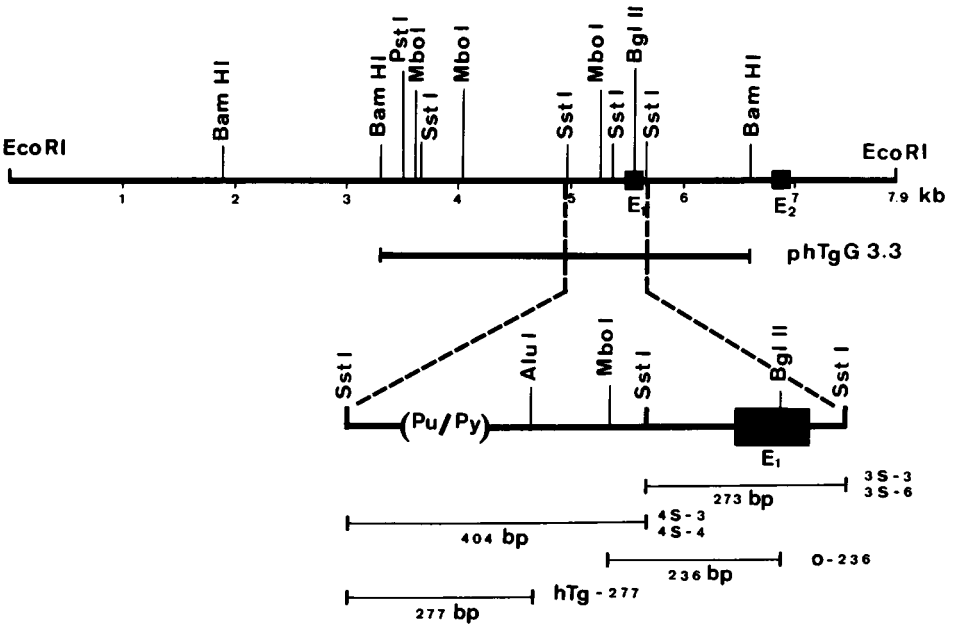


Fig. 1. Restriction map of the 5' end of the human Tg gene showing the location of the two first exons (black boxes) and the cloned sequences used in this study. For clarity only those restriction sites which are relevant to the work have been indicated.

cloned in plasmid pBR322 (phTgG 3.3; see fig. 1). Taking in account the electron microscopy data obtained previously (18) and the knowledge of the exonic sequence (19), restriction mapping experiments lead to the localization of the first exon in a small, 273 bp-long Sst I fragment. This was inserted in M 13 mp 10 phage DNA in both orientations and cloned (couple of clones 3S-3/3S-6; fig. 1). The upstream contiguous Sst I fragment was also subcloned (couple of clones 4S-3/4S-4, fig. 1). These four M13 clones allowed us to determine the sequence displayed in Fig. 2. In order to confirm the sequence at the junction of the two Sst I fragments, we subcloned and sequenced the 236 bp-long Sau 3A fragment overlapping the Sst I site (clone 0-236; fig. 1).

Identification of the first exon of the gene. Since the cDNA sequence of human Tg was known (19), the 3' end of the first exon was readily identified by direct comparison of the sequences. The exon-intron junction splits the codon corresponding to the glutamic acid residue which is the fourth amino-acid of the mature protein. This amino-acid is also the one preceeding the first tyrosine residue in Tg sequence, which has been shown to be converted

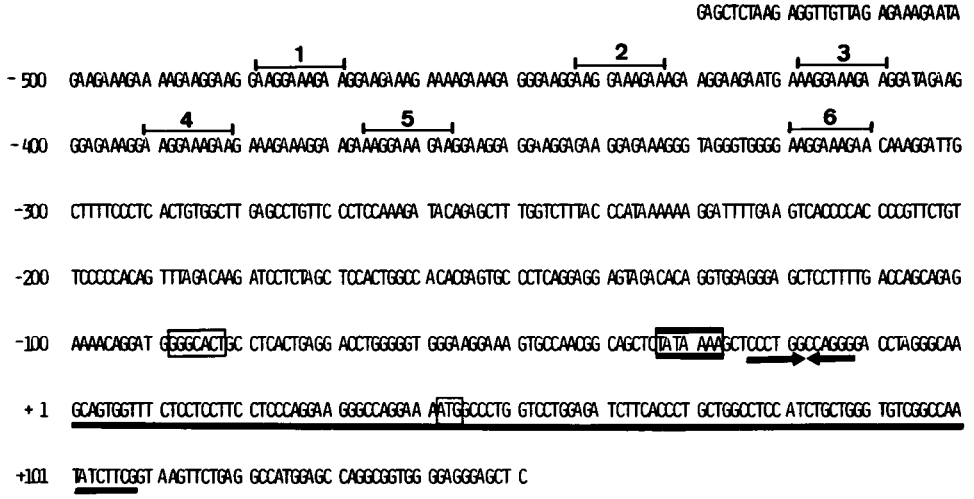


Fig. 2. Sequence of the 5' end of the human Tg gene. Position +1 corresponds to the transcriptional start site. The CAAT-box homology, the TATA-box and the translation initiator codon are boxed. The sequence of the first exon is underlined. The 12 bp palindromic sequence is indicated by the two head-to-head arrows. The repeated motifs contained in the homopurine sequence are over-lined and numbered from 1 to 6.

into thyroxine during hormonogenesis (4). The DNA sequence corresponding to the amino-terminal end of the mature Tg is preceeded by a region encoding a 19 amino-acids long signal peptide. The translation-initiator ATG complies to Kozak's rule (20) in being preceeded by a A in position -3. Our sequence is in perfect agreement with that described by Malthiery et al (19) as far as the translated part of the mRNA is concerned. However, many discordances are found upstream from the translation-initiator codon. Even if the occurence of a polymorphism in the 5' non-translated sequence cannot be excluded, it seems most likely that this discrepancy results from the use of different sequencing strategies. As only our sequence has been obtained by sequencing both strands of the DNA template we give it the preference.

To localize precisely the transcriptional start-site, a chemically synthesized 16-mer primer ($^{5'}$CAGGGTGAAGATCTCC$^{3'}$) complementary to a sequence contained in the first exon was 5'-labeled, hybridized to human Tg mRNA and extended in a reverse transcription reaction. The lengths of the resulting cDNAs were measured on a 8% acrylamide sequencing gel (fig. 3). Two major cDNAs species were detected, 71 bp- and 43 bp-long respectively. Considering the size of the longest cDNA we deduced that transcription initiates at the G residue

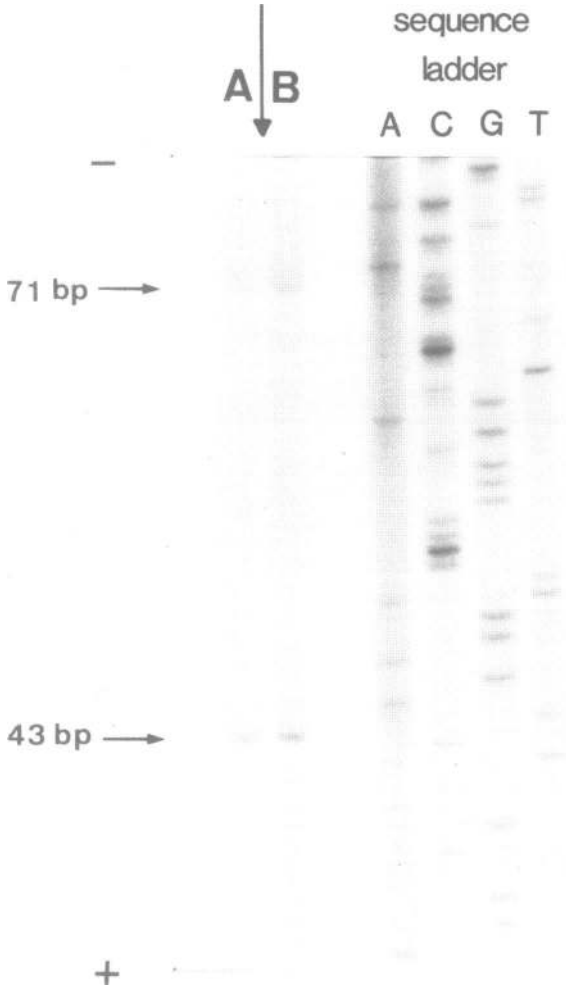


Fig. 3. Localization of the transcriptional start site by primer extension. Autoradiography of the 8 % acrylamide-sequencing gel. The cDNA products obtained from two different RNA preparations from normal human thyroid tissue were run on lanes A and B respectively. (The sequence ladder used as size marker was obtained from an unrelated M13 clone).

("+1", see fig. 2) located 28 bp downstream from a canonical TATA Box (TATAAAA). The shorter cDNA could then correspond to a second transcriptional start site, 28 bp downstream from the former, or result from unspecific priming or from a (quite specific) premature termination of Tg mRNA reverse transcription. Heterogeneity in transcription initiation, when present, is mostly limited to a very small region around a major transcriptional start-

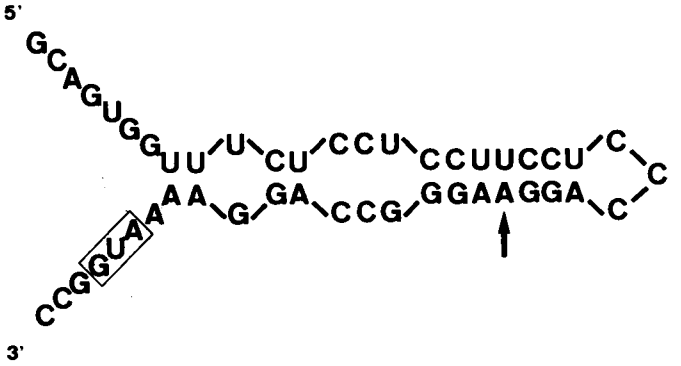


Fig. 4. Possible hairpin structure of the 5' non-translated sequence of human Tg mRNA.
The translation-initiator AUG is boxed; the arrow indicates the position corresponding to the 5' end of the 43 bp cDNA product (see text).

site (e.g. ovalbumin gene (21,22)). A more striking heterogeneity (30 bp) has been reported in the case of the α -amylase gene transcription in the mouse liver (23). This was further demonstrated to result from the use of an overlapping promoter containing a degenerated TATA Box (24). No such cryptic motif can be found in front of the Tg gene. Also, according to the "TATA Box-trap" model of Bensimhon et al (25), a TATA motif as closely flanked by GC-rich sequence as that of the Tg promoter, should direct the initiation of transcription very sharply. Therefore the existence of a second transcriptional start site remains speculative. In contrast, several possibilities exist which could lead to the generation of an artefactual band. No priming of cDNA synthesis occured when the labeled primer was hybridized to total RNA extracted from canine sertolioma (not shown). Although this argues in favor of specificity of priming in our conditions, it does not rule out the possibility that RNA sequences harboring a fortuitous homology to the primer are present in total RNA from human thyroid. Also, since priming of the reverse transcription can occur on hybrids only a few base-pairs long, especially in GC-rich region (26), wrong priming could take place and result in the synthesis of an artefactual cDNA species in addition to the true extension product. Finally the possibility of forming a secondary structure at the 5' end of the mRNA could still provide another explanation (see below and fig. 4).

Considering the 5' non-translated sequence of the mRNA as defined by the canonical cap site (see fig. 2), a 33 bp quasi-palindromic sequence is observed which could possibly adopt a hairpin structure on folding (11 matches, see fig. 4). Similar structures have been described in the 5' untranslated region

of the collagen genes (27). Their possible involvement in translational control has been hypothesized.

The promoter region contains a long homopurine-homopyrimidine stretch

Apart from the already cited TATA motif, the promoter exhibited a CAAT-Box homology (GGGCACT) located -89 to -83 bp relative to the cap site. A 12 bp-long GC-rich palindromic sequence was also found between the TATA-Box and the transcriptional start site.

No significant sequence homology was found when the Tg gene promoter was compared with the corresponding region of other genes the transcription of which is known to be regulated by cAMP (12,13,14). Likewise, there is no indication that an enhancer element structurally related to either the SV-40 or the adeno-virus type (28) would lie in this region.

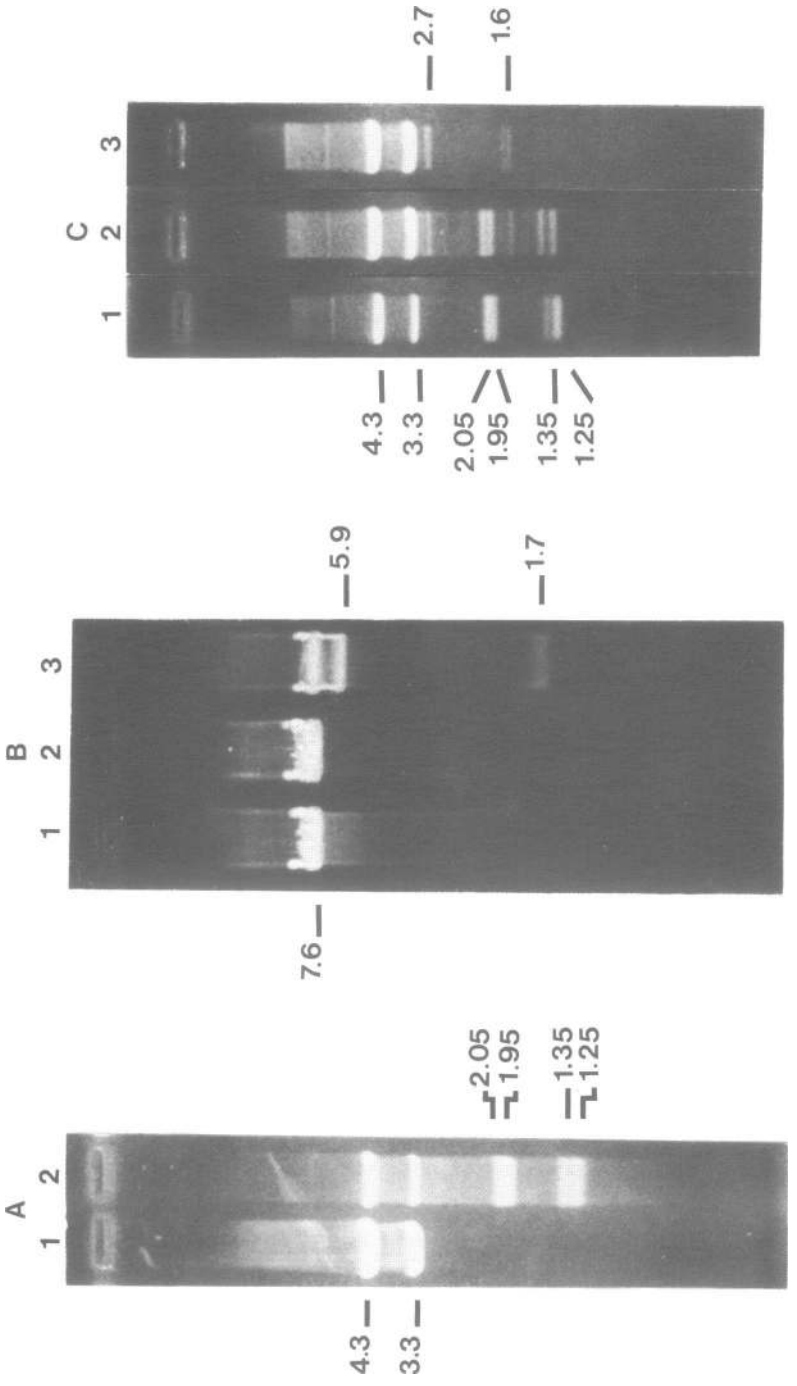
The prominent feature was obviously the 209 bp-long homopurine-homopyrimidine sequence located between positions -512 to -304. Only five isolated T and one C residues interrupt the homopurine stretch which is on the coding strand. This segment exhibits a rather complicated sequence organization. It presents an overall homology of 74 % with the sequence $(AGGA)_{52}$. However, only the first 160 bases account for this high score since they display a 82,5% homology with the sequence $(AGGA)_{40}$. Beyond this point the homology drops to 58 %, a value which has no statistical significance in a sequence made of only two base species. By contrast, a 10 bp-long motif (AAGGAAGAA) is found repeated six times throughout the sequence. Two couples of these repeated blocs (2-3 and 4-5, see fig 2) are separated by the same number of bases. This repeated pattern of a unique motif diverging from the monotonous consensus $(AGGA)_n$ strongly suggest that this sequence originated from sequential duplication events.

The homopurine-homopyrimidine sequence is cut by S_1-nuclease when contained in a supercoiled plasmid.

Similar homopurine-homopyrimidine sequences have been found in the vicinity of a number of genes, mostly in the 5' flanking region (29-32). In somes cases, they have also been shown to be repeated several times throughout the genome (33, 34). When tested, such segments exhibited sensitivity to the single-strand specific S_1 endonuclease when they were subcloned in a supercoiled plasmid. We therefore tested the sensitivity to S_1 nuclease of the plasmids containing the homopurine-homopyrimidine sequence present in the Tg promoter region. In conditions where no significant degradation of supercoiled pBR322 plasmid was detectable, the enzyme specifically cuts the supercoiled recombi-

nants in the homopurine-homopyrimidine sequence, as deduced from subsequent restriction analysis (fig. 5). Moreover, when S_1 nuclease-treated phTgG3.3 plasmid DNA was restricted with Bam HI, two closely related populations of fragments were detected after gel electrophoresis (fig. 5a). This indicated that S_1 nuclease cuts the sequence at two distinct positions the most likely explanation being that one digestion event excluded the other by relaxing the superhelicity of the substrate . This reaction was indeed clearly supercoil-dependent since prior linearization of the plasmid by EcoRI totally impaired subsequent cleavage by S_1 nuclease (fig. 5b). The distance between the two cutting sites was estimated to be about 100 bp. The same digestion pattern was observed when Mung Bean Nuclease was used instead of S_1 nuclease. However, raising the pH of the medium from pH 4.5 to pH 6 lead to a gradual suppression of the reaction (fig. 5c) (the ability of the enzyme to function effectively in these conditions was assayed in parallel on single-stranded DNA). The pH related disappearance of the sensitivity in the homopurine-homopyrimidine region unmasked cleavage at minor sites in the pBR322 portion of the plasmid, resulting in two faint bands 1.6 and 2.7 kb in length. These fragments correspond to cleavage at a known S_1 nuclease-sensitive site in supercoiled pBR322 (35).

The supercoiled replicative forms of the M13 recombinants containing the homopurine-homopyrimidine sequence also exhibited sensitivity towards S_1 nuclease action. This was exploited to map more precisely the positions where the enzyme cuts each DNA strand. An M13 clone which contains the homopurine-homopyrimidine sequence essentially was constructed for this purpose (hTg-277, see fig. 1). We first checked that the isolated homopurine-homopyrimidine sequence still exhibited the same cleavage pattern as that obtained with the longer insert. This was achieved by running on a 6%-acrylamide gel the fragments obtained after S_1 nuclease digestion and insert excision from the various clones. The results (not shown) indicated that the sequence was cut at the same positions in all cases and allowed us to localize approximatively these two positions relative to the Alu I restric tion site used for the construction of clone hTg 277 (see fig. 6c). Taking advantage of the existence of sequencing primers for both strands of M13 DNA we characterized the cutting sites at the base level by extending these primers on S_1 nuclease-cleaved and heat-denatured hTg-277 RF DNA (fig. 6a). This allowed us to localize on the sequence the 5' borders of the nicks created by the enzyme in the homopurine-homopyrimidine region (fig. 6b). As already observed by others (29, 31, 32, 36), the two strands of the DNA were not cut in a symme-


A
1
2
4.3
3.3
2.05
1.95
1.35
1.25
B
1
2
3
7.6
5.9
1.7
C
1
2
3
4.3
3.3
2.05
1.95
1.35
1.25
2.7
1.6

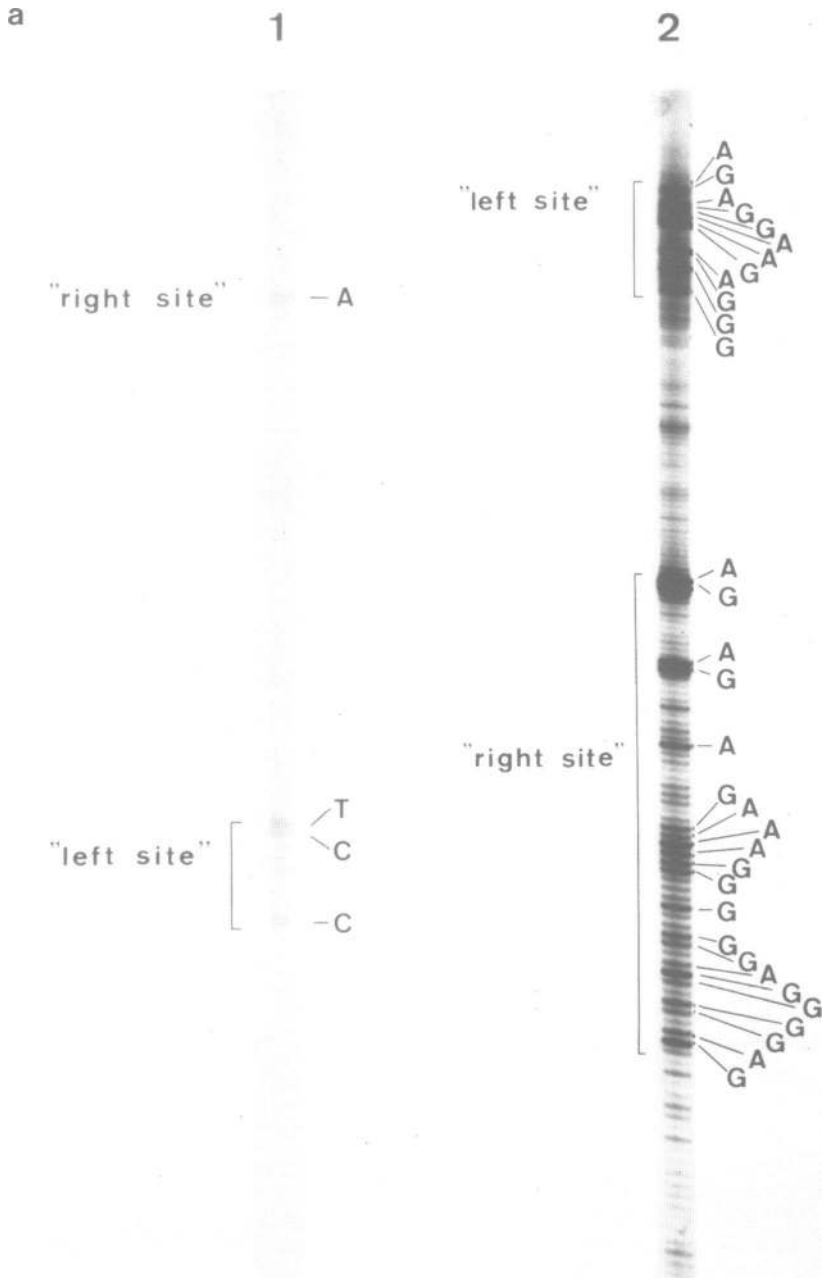
trical way. The extension products derived from the pyrimidine strand of the template showed a very narrow range of lengths, reflecting a highly specific action of the enzyme. In contrast, those obtained from the purine template were quite heterologous in size. Considering together the data obtained on both strands, a clear correspondance appeared between the position of the two clusters of cutting sites and the size of the fragments generated by S1 digestion followed by restriction (fig. 6c).

DISCUSSION.

The 5' region of the human Tg gene has been characterized. When compared to genes encoding other secreted peptides, the first exon displays a classical composition : a 40 bp non-translated sequence preceeds an open-reading frame encoding the 19 amino-acids long leader peptide, fused to the mature protein coding sequence. The untranslated sequence contains a quasi-palindromic region which could adopt a hairpin structure. Another peculiar feature is found in the promoter region of the gene which, apart from the canonical TATA Box and a CAAT Box homology, also contains a 209 bp-long homopurine-homopyrimidine sequence. This sequence exhibits a dual character being both monotonous, as it presents an overall 74 % homology with the sequence $(AGGA)_{52}$, as well as structured since it displays an organized pattern of six 10 bp-long repetitive units (fig. 2).

This homopurine-homopyrimidine sequence is cut by single-strand specific endonucleases when contained in supercoiled molecules. The reaction is supercoil - and pH - dependent. The cleavage can occur at two distinct positions in the homopurine-homopyrimidine region, but, due to the supercoiling requirement, each molecule can be cut only once. Despite a growing number of studies (36, 37, 38, 39) the reason why homopurine-homopyrimidine sequences exhibit sensitivity towards single-strand specific endonuclease under torsionnal stress remains poorly understood. Models involving a slippage mechanism in regions containing tandem repeats (29,40) or non-Z left-

Fig. 5. Sensitivity of supercoiled phTgG3.3 plasmid DNA to S_1 nuclease and Mung Bean nuclease.
Fragments lengths are given in kb.
A) phTgG3.3 DNA restricted by Bam HI whitout (lane 1) or after (lane 2) previous S_1 nuclease treatment.
B) phTgG3.3 DNA restricted by Eco RI before (lane 1), whitout (lane 2) or after (lane 3) S_1 nuclease treatment.
c) phTgG3.3 DNA restricted by Bam HI after Mung Bean nuclease treatment at pH 4.5 (lane 1) , pH 5 (lane 2) or pH 6 (lane 3).


a
1
2
"left site"
"right site"
A
"right site"
T
C
"left site"
C

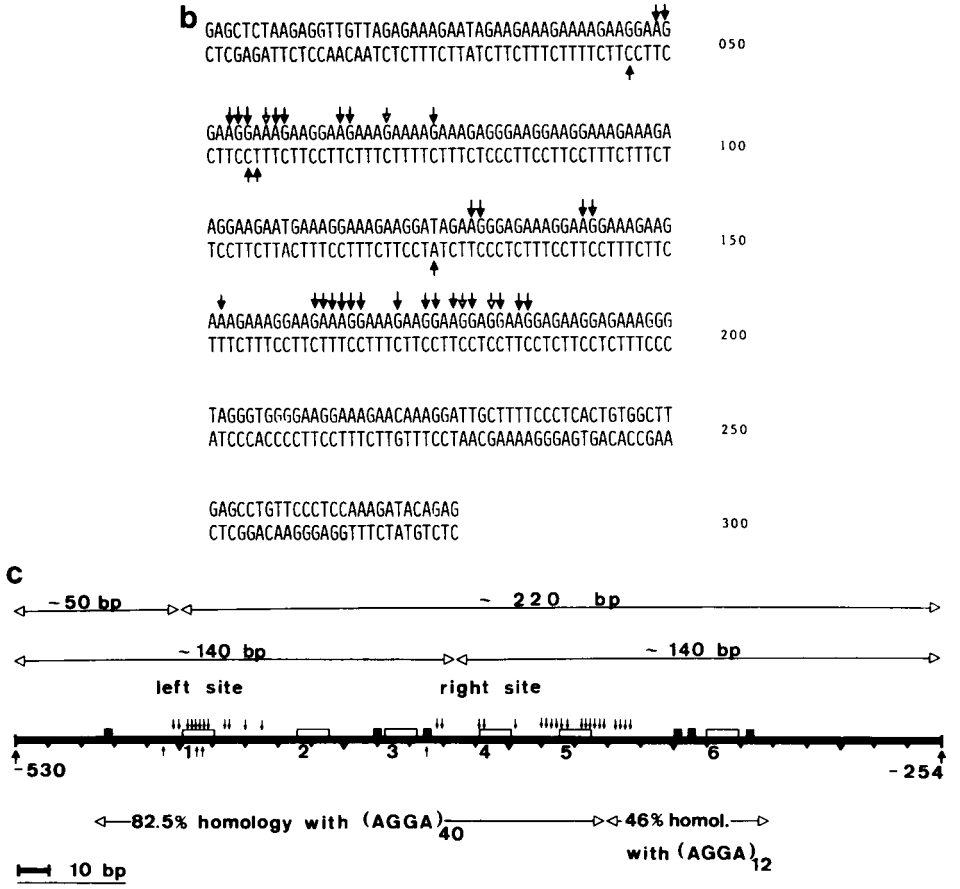


Fig. 6. Mapping of S_1 nuclease-sensitive sites in the 5'-flanking region of the human Tg gene.

a. Autoradiograms of sequencing gels showing the products resulting from extension of M13 sequencing (lane 1) and reverse-sequencing (lane 2) primers hybridized to S_1 nuclease-treated hTg-277 DNA. The correspondence to template bases is given for the major extension products.

b. Double-stranded sequence of the insert contained in clone hTg-277. The arrows indicate the extremities of the major extension products.

c. Schematical representation of the cleavage pattern of the homopurine-homopyrimidine sequence by S_1 nuclease. The horizontal line represents the insert sequence contained in clone hTg-277 (coordinates refer to Tg gene Cap site; see fig. 2). The black squares correspond to base inversions in respect to the homopurine-homopyrimidine asymmetry, and the open boxes symbolize the repeated units. The arrows localize the 5' borders of the gaps on the purine (top) and pyrimidine (bottom) strands. The average lengths of the fragments detected by gel electrophoresis after S_1 nuclease treatment and excision of the insert are indicated.

handed DNA structures (39) seems most favoured. In our case the cleavage positions do not correlate well with the presence of multiple tandem repeats. However, a slippage mechanism cannot be definitely excluded, especially since the gaps in the purine strand are mainly detected at the level of repeated units 1 (left site) and 5 (right site) (fig. 2). Another possibility could be that the slippage would involve the quasi-monotonous sequence non specifically. In this case however, one would not expect finding such precisely located cutting sites. The occurence of an altered non-B non-Z DNA structure cannot be ruled out neither. However, we propose another explanation which agrees with all our observations concerning the behaviour of this homopurine-homopyrimidine sequence. We suggest that the observed S_1 nuclease sensitivity results from the local formation of a triple-helix DNA structure in the homopurine-homopyrimidine region at low pH. It has been shown that $d(TC)_n.d(GA)_n$ duplex DNA forms triple - helical complexes at low pH (41). These complexes appear to form spontaneously at neutral pH when all cytosine residues are methylated (42). In our case, the 160 bp-long region of the homopurine-homopyrimidine sequence which shows an overall monotonous character (82.5%) homology with $(AGGA)_{40}$) could adopt such kind of structure at low pH upon folding-back under torsionnal stress. Once the structure is formed it should be very stable in acidic conditions (41). In the triple-helix structure, base-triads are formed between duplex DNA and a protonated pyrimidine strand. The corresponding purine sequence is left unpaired. Therefore, two equivalent or symmetrical structures can be formed by folding back the sequence. This could account for the presence of two cutting sites (see fig. 7). According to this model, one can predict that S_1 nuclease would cut the pyrimidine strand quite sharply at both extremities of the triple-helix structure. In contrast, the portion of the purine strand which is excluded from the triplex could be digested randomly. If considering that the strategy used in this study allowed the mapping of the 5' borders of the nicks only, the pattern of strand-breaks we observed fully agrees with these predictions. When compared to the single-base cut at the right site, the cleavage on the pyrimidine strand at left site appears more heterogenous. This could be explained by differences in the extent of triple-helix stabilization by supercoiling. We should point out that such a kind of model represents an alternative to the slippage-mechanism hypothesis (29, 40) as long as the distance between the repeats still allows the formation of a double-stranded hairpin. This condition is fullfilled, for example, in the case of the S_1 nuclease- sensitive homopurine-homopyrimidine sequence located in the human

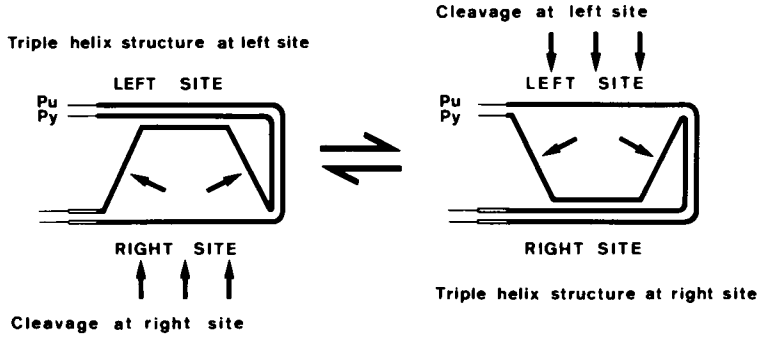


Fig. 7. Schematic representation of the proposed model for S_1 nuclease cleavage in the homopurine-homopyrimidine region.

——— flanking sequences

▬▬ homopurine-homopyrimidine region homologous to $(AGGA)_n$

▭▭ homopurine-homopyrimidine region non-homologous to $(AGGA)_n$

The arrows point to S_1 nuclease-sensitive (single-stranded) regions. (The drawings have been distended intentionally for clarity; the single-stranded portions of the pyrimidine strand should be a few bases long only).

α_2- α_1 globin intergenic region (43). The distribution of cutting sites observed in this case after labeling of the 3' extremities of the nick is compatible with the occurence of a hairpin-triplex structure. This would involve a region essentially made of the imperfect repetition of a CCT motif.

What could be the role, if any, of these homopurine-homopyrimidine sequences ? To date, there seems to be no satisfactory answer. Some observations support the view that they could be involved in the control of gene expression. Larsen and Weintraub (44) first demonstrated that the presence of S_1 nuclease-sensitives regions in chicken globin chromatin is associated with gene activity. The sequences involved in the sensitivity at the chromatin level were shown to be the same as those cleaved by the enzyme in supercoiled recombinant plasmids. An S_1 nuclease-sensitive homopurine-homopyrimidine stretch is also conserved in the promoter region of the chicken and mouse α_2 (I) collagen genes, despite extensive divergence in neighbouring sequences (40). Preliminary results from transfection experiments suggest that this sequence could be required for the efficient transcription of the gene. In both globin and collagen genes, the S_1 nuclease-sensitive sequences map in DNAse I-hypersensitive regions of the chromatin. As these domains are thought to be the target for protein-DNA interactions relevant to gene expression

(45), the nearby presence of the homopurine-homopyrimidine sequences argues for their participation in the regulation of gene activity.

A model for the control of gene expression involving the formation of a triple-helix structure in homopurine-homopyrimidine regions has been recently proposed (42). In order to occur under physiological conditions, triplex formation would require a stabilizing mechanism mimicking the cytosine-protonation effect. A protein-DNA interaction or cytosine methylation could conceivably play this role. Although 5-methyl- cytosine is mostly found in CG sequences (46), it cannot be excluded (47) that a few homopyrimidine sequences would be highly methylated on C residues. The hypothesis that tissue-specific methylation at homopyrimidine sequences would correlate with gene expression is supported by the observation that the homopyrimidine regions located up stream of genes characterized to date always contain C residues (30, 32, 33, 40, 43). Methylation at these sequences would not have been detected by restriction analysis due to their peculiar base composition. Verification of this hypothesis represents therefore an examplative application for the use of the recently described genomic-sequencing technique (48). On the other hand, protein-DNA interactions at homopurine-homopyrimidine sequences (40) or at regions containing homopurine-homopyrimidine stretches (49) have been recently reported. In this context, it is noteworthy that the formation of a hairpin structure in the 5' untranslated region of Tg mRNA would only involve homopurine-homopyrimidine base pairs. In addition to the well established transcriptional control (7), Tg synthesis in the thyroid appears to be also controlled by TSH at the translational level (50). It is therefore tempting to hypothesize that both levels of regulation would involve similar sequence recognition mechanisms.

ACKNOWLEDGMENTS

The continuous support and critical interest of Professor J.E. Dumont is deeply acknowledged. We express our gratitude to Dr A. Cravador for the synthesis of the primer complementary to the 5' end of the mRNA, and to Dr S. Swillens for the elaboration of the computer programs used in the analysis of our nucleic-acids sequences. We thank Mrs D. Leemans for the preparation of the manuscript.
This study has been supported by grants from the Ministère de la Politique Scientifique (Action Concertée) to J.E. Dumont, from NIH (n° AM 21732) and FRSM (n° 3433479) to G.V. and from Association Recherche Biomedicale et Diagnostic asbl. A.B. is a research fellow of the E.E.C.. B.C. and H.T. were recipients of fellowships from EMBO and CONICET (Argentina) respectively.

REFERENCES

1. van Ommen, G.-J.B., Arnberg, A.C., Baas, F., Brocas, H., Sterk, A., Tegelaers, W.H.H., Vassart, G. and de Vijlder, J.J.M. (1983) Nucl. Acids Res. 11, 2273-2285.
2. Vassart, G. and Brocas, H. (1980) Biochim. Biophys. Acta 610,189-194
3. Marricq, C., Arnaud, C., Rolland, M. and Lissitzky, S. (1980) Eur. J. Biochem. 111, 33-47.
4. Mercken, L., Simons, M.-J. and Vassart, G. (1982) FEBS Lett. 149, 285-287.
5. Mercken, L., Massaer, M., Simons, M.-J., Swillens, S. and Vassart, G. (1984) Biochem. Biophys. Res. Commun. 125, 961-966.
6. Van Herle, A.J., Vassart, G. and Dumont, J.E. (1979) New Engl. J. Med. 301, 239-249 and 307-314.
7. Van Heuverswyn, B., Streydio, C., Brocas, H., Refetoff, S., Dumont, J. and Vassart, G. (1984) Proc. Natl. Acad. Sci. USA 81, 5941-5945.
8. VanSande, J., Cochaux, P., Mockel, J. and Dumont, J.E. (1983) Mol. Cell. Endoc. 29, 109-119.
9. Van Heuverswyn, B., Leriche, A. and Vassart, G. (1984) Ann. Endoc. 45, 10.
10. Groner, B., Ponta, H., Beato, M. and Hynes, N.E. (1983) Mol. Cell. Endoc. 32, 101-116.
11. Parker, M.G. and Page, M.J. (1984) Mol. Cell. Endoc. 34, 159-168.
12. Wynshaw-Boris, A., Lugo, T.G., Short, J.M., Fournier, R.E.K. and Hanson, R.W. (1984) J. Biol. Chem. 259, 12161-12169.
13. Shinomiya, T., Scherer, G., Schmid, W., Zentgraf, H. and Schutz, G. (1984) Proc. Natl. Acad. Sci. USA 81, 1346-1350.
14. Truong, A.T., Duez, C., Belayew, A., Renard, A., Pictet, R., Bell, G.I. and Martial, J.A. (1984) EMBO, J. 3, 429-437.
15. Maniatis, T., Fritsch, E.F. and Sambrook, J. Molecular Cloning, a laboratory manual (1982) Cold Spring Harbor Laboratory, New York.
16. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) Proc. Natl. Acad. Sci. USA 74, 5463-5467.
17. Aufray, C. and Rougeon, F. (1980) Eur. J. Biochem. 107, 303-314.
18. Targovnik, H.M., Pohl, V., Christophe, D., Cabrer, B., Brocas, H. and Vassart, G. (1984) Eur. J. Biochem. 141, 271-277.
19. Malthiery, Y. and Lissitzky, S. (1985) Eur. J. Biochem. 147, 53-58.
20. Kozak, M. (1984) Nucl. Acids Res. 12, 857-872.
21. Malek, L.T., Eschenfeldt, W.H., Munns, T.W. and Rhoads, R.E. (1981) Nucl. Acids. Res. 9, 1669-1673.
22. Heilig, R., Murakowsky, R. and Mandel, J.-L. (1982) J. Mol. Biol. 156, 1-19.
23. Hagenbuchle, O., Tosi, M., Schibler, U., Bovey, R., Wellauer, P.K. and Young, R.A. (1981) Nature 289, 643-646.
24. Schibler, U., Pittet, A.-C., Young, R.A., Hagenbuchle, O., Tosi, M., Gellman, S. anf Wellauer, P. (1982) J. Mol. Biol. 155, 247-266.
25. Bensimhon, M., Gabarro-Arpa, J., Ehrlich, R. and Reiss, C. (1983) Nucl. Acids Res. 11, 4521-4540.
26. Fukui, T., Ishihama, A., Ohtsuka, E., Ikehara, M. and Fukuda, P. (1982) J. Biochem. 91, 331-339.
27. Yamada, Y., Mudryj, M. and de Crombrugghe, B. (1983) J. Biol. Chem. 258, 14914-14919.
28. Yaniv, M. (1984) Biol. Cell. 50, 203-216.
29. Mace H.A.F., Pelham, H.R.B. and Travers, A.A. (1983) Nature 304, 555-557.
30. Nickol, J.M. and Felsenfeld, G. (1983) Cell 35, 467-477.
31. Finer, M.H., Fodor, E.J.B., Boedtker, H. and Doty, P. (1984) Proc. Natl. Acad. Sci. USA 81, 1659-1663.

32. Ruiz-Carrillo, A. (1984) Nucl. Acids. Res. 12, 6473-6492.
33. Maroteaux, L., Heilig, R., Dupret, D. and Mandel, J.L. (1983) Nucl. Acids Res. 11, 1227-1243.
34. Dybvig, K., Clark, C.D., Aliperti, G. and Schlesinger, M.J. (1983) Nucl. Acids. Res. 11, 8495-8508.
35. Lilley, D.M.J. (1980) Proc. Natl. Acad. Sci. USA 77, 6468-6472
36. Schon, E., Evans, T., Welsh, J. and Efstratiadis, A. (1983) Cell 35, 837-848.
37. Weintraub, H. (1983) Cell 32, 1191-1203.
38. Evans, T., Schon, E., Gora-Maslak, G., Patterson, J. and Efstratiadis, A. (1984) Nucl. Acids Res. 12, 8043-8058.
39. Cantor, C.R. and Efstratiadis, A. (1984) Nucl. Acids Res. 12, 8059-8072.
40. McKeon, C., Schmidt, A. and de Crombrugghe, B. (1984) J. Biol. Chem 259, 6636-6640.
41. Lee, J.S., Johnson, D.A. and Morgan, A.R. (1979) Nucl. Acids Res. 6, 3073-3091.
42. Lee, J.S., Woodsworth, M.L., Latimer, L.J.P. and Morgan, A.R. (1984) Nucl. Acids Res. 12, 6603-6614.
43. Shen, C.-K.J. (1983) Nucl. Acids. Res. 11, 7899-7910.
44. Larsen, A. and Weintraub, H. (1982) Cell 29, 609-622.
45. Elgin, S.C.R. (1984) Nature 309, 213-214.
46. Razin, A. and Szyf, M. (1984) Biochim. Biophys. Acta 782, 331-342.
47. Grippo, P., Iaccarino, M., Parisi, E. and Scarano, E. (1968) J. Mol. Biol. 36, 195-208.
48. Church, G.M. and Gilbert, W. (1984) Proc. Natl. Acad. Sci. USA 81, 1991-1995.
49. Dilworth, S.M., Cowie, A., Kamen, R.I. and Griffin, B.E. (1984) Proc. Natl. Acad. Sci. USA 81, 1941-1945.
50. Davies, E., Dumont, J.E. and Vassart, G. (1978) Biochem. J. 172, 227-231.