Proc. Natl. Acad. Sci. USA
Vol. 83, pp. 323–327, January 1986
Biochemistry

The complete structure of the rat thyroglobulin gene

(thyroxine/thyroid/duplication/shuffling)

ANNA M. MUSTI, ENRICO V. AVVEDIMENTO, CLAUDIO POLISTINA, VALERIA M. URSINI, SILVANA OBICI, LUCIO NITSCH, SERGIO COCOZZA, AND ROBERTO DI LAURO*†

Centro di Endocrinologia ed Oncologia Sperimentale del Consiglio Nazionale della Ricerche, c/o Dipartimento di Biologia e Patologia Cellulare e Molecolare, II Facolta' di Medicina e Chirurgia, Via S. Pansini 5, 80131 Napoli, Italy

Communicated by Maxine F. Singer, September 19, 1985

ABSTRACT **We have isolated the entire gene for rat thyroglobulin, the precursor for thyroid hormone biosynthesis. The gene is at least 170,000 base pairs (bp) long; 9000 bp of coding information are distributed in 42 exons of homogeneous size (150–200 bp) except for two exons of 1100 and 620 bp. The sequences coding for two major thyroxine-forming sites are localized in exons 2 and 39. These two sequences do not show any homology either at the DNA or at the protein-sequence level, even though they code for sites highly specialized for the same function. Furthermore, both the 3′ and the 5′ end of the thyroglobulin structural gene appear to be made of repetitive units, which again do not show any homology. On the basis of these observations, we propose that the thyroglobulin gene arose by shuffling of at least two segments, with different evolutionary histories, each of which already contained introns.**

The two hormones secreted by the thyroid gland, thyroxine (T4) and 3-3′-5-triiodothyronine (T3), are synthesized by the coupling of a diiodinated tyrosyl residue with either another diiodinated tyrosyl residue or with a monoiodinated tyrosyl residue, respectively. The coupling occurs within the polypeptide chain of thyroglobulin (Tg), a very large thyroid-specific glycoprotein made of two identical subunits (1). In spite of such a size, only 3–5 mol of hormone are formed per mol of protein (2, 3). The primary structure of two sites where most of the thyroxine is formed (the hormonogenic sites) has been determined by direct protein sequencing (4, 5), whereas most of the information on the primary structure of the protein has been deduced from the nucleotide sequence of portions of cDNA for human (6), bovine (7), and rat (8) Tg. The analysis of the protein and of the mRNA sequence has shown that the two major hormonogenic tyrosines are located at the extremities of the protein (6–9). This paper presents an overall view of an entire Tg gene, the rat gene.

The rat gene encoding Tg is made of at least 170,000 base pairs (bp), about 20 times larger than its corresponding mRNA (8). The coding information is divided into 42 exons of homogeneous size (150–200 bp) except for two exons of 1100 and 620 bp at the 5′ end. Several reports have already appeared on segments of the Tg gene from different species (10–13). We have suggested before (8, 14), on the basis of interspecies sequence comparison, that the 3′ and the 5′ end of the Tg gene may have two independent evolutionary origins. To analyze this hypothesis further, it is important to compare the two ends of a gene from one species. In this paper we analyze the nucleotide sequence of the exons and the sequences coding for the two hormonogenic sites at the two ends of the rat gene. We find that there is no homology between the two thyroxine-forming sites in rat Tg. Furthermore, the two sites are located in gene segments that have a completely unrelated sequence organization, as shown by the presence of two different repeated units.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

MATERIALS AND METHODS

Libraries and Vectors. Two rat genomic libraries have been used in this study: the *Hae* III/*Alu* I partial-digest library was kindly provided by J. Bonner (15); the *Eco*RI partial-digest library was constructed in our laboratory as described (12). Screenings of libraries, purification of λ phage DNA, and probe labeling by nick-translation were carried out by published protocols (16). R loops were obtained and analyzed as described by Davis *et al.* (17). Repetitive elements were mapped in the recombinant λ phages (12) and DNA was sequenced (18) as described.

RESULTS

cDNA Clones Used in the Isolation of the Rat Tg Gene. We have described (8) a set of cDNA clones derived from the rat Tg mRNA. On the basis of a primer elongation experiment, we suggested that the cDNA clone pRT27.15 extended up to the 5′ end of the mRNA.

We subsequently discovered, comparing our partial sequence data with the sequence of the Tg mRNA in humans (6) and cows (7), that clone pRT27.15 stops at about 1 kilobase (kb) from the 5′ end of Tg mRNA. We then isolated other cDNA clones (data not shown), and the one extending the most at the 5′ end is clone pRT27.23, which ends at 450 nucleotides from the 5′ end of the mRNA (Fig. 1*A*).

Physical Map of the 5′ Region of the Rat Tg Gene. The new collection of cDNA clones was used to complete the isolation of the rat Tg gene. Six recombinant λ phages (2, 43, 49, 16, 18, 106) were obtained from the screening of two different rat genomic libraries (see the legend of Fig. 1). The relative position and the 5′–3′ polarity of the isolated genomic segments was established by hybridization with appropriate restriction fragments derived from the cDNA clones (Fig. 1). All of the genomic clones shown in Fig. 1 are overlapping, except λ49 and λ16, which nevertheless contain continguous exons, as demonstrated by the hybridization of the cDNA fragment c (Fig. 1) with digested rat DNA. All of the genomic fragments detected by fragment c are present in phages λ16 and λ49, suggesting that only an intronic segment bridges the two recombinant clones (data not shown). The two overlapping phages at the 5′ end of the gene (λ10 and λ63) were isolated by using as a probe the fragment G at the 5′ end of phage λ106 (Fig. 1). In order to prove that the 5′ end of the rat Tg gene is contained in the phages λ10 and λ63, restriction

Abbreviations: Tg, thyroglobulin; bp, base pair(s); kb, kilobase(s).

*To whom correspondence should be sent.
†Present address: Laboratory of Biochemistry, Building 37, Room 4C-13, National Cancer Institute, Bethesda, MD 20892.

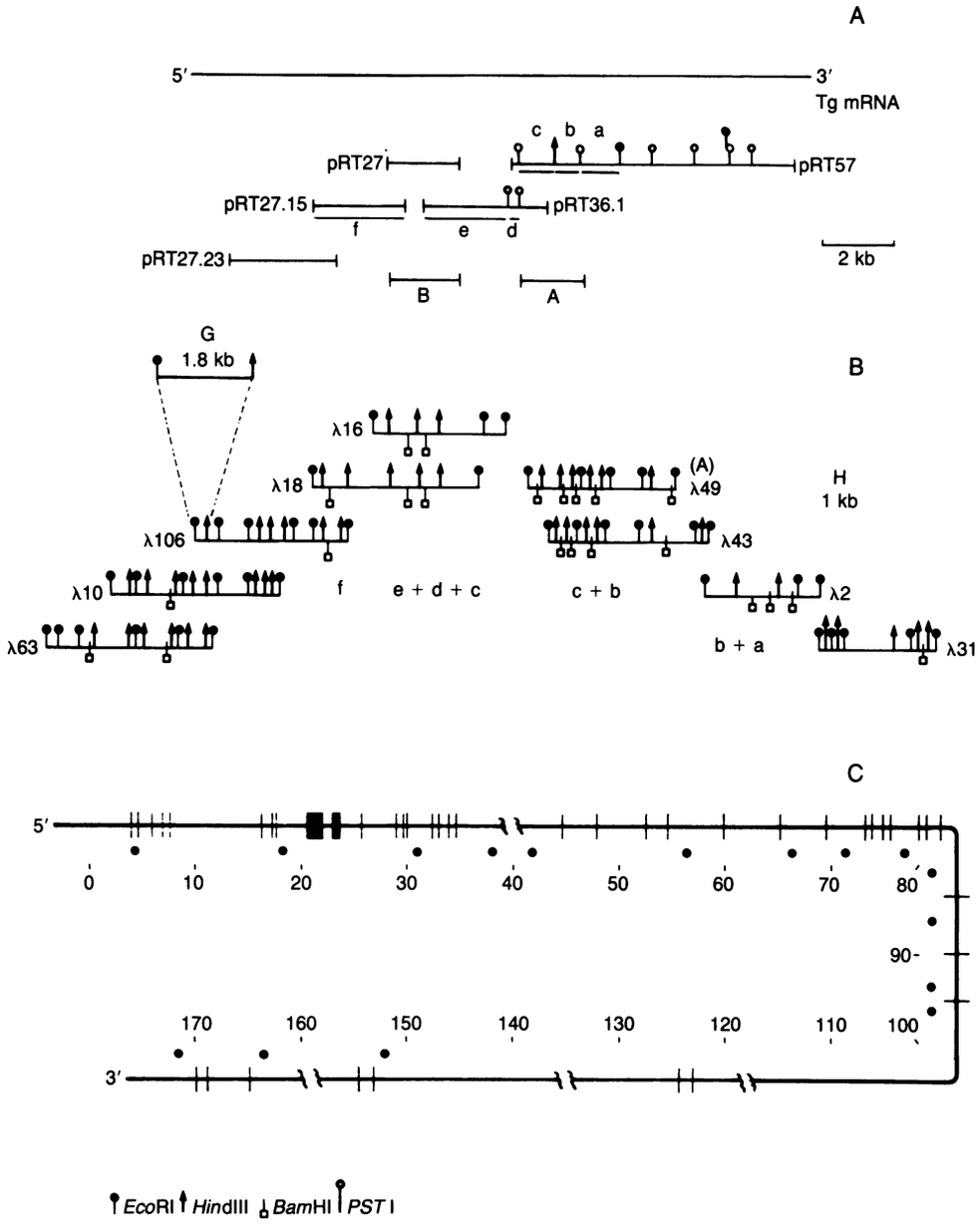

FIG. 1. Physical map of the rat Tg gene. (*A*) cDNA clones for the rat Tg mRNA. The cDNA fragments A and B represent the probes used for the screening of the genomic libraries. The cDNA fragments a, b, c, d, e, and f represent the probes used to determine the polarity of the recombinant phages. (*B*) Restriction map of the recombinant phages containing the 5′ half of the rat Tg gene. Phage λ31 has been reported (11). The phages have been drawn so that their position in the gene can be deduced by projecting them on the top line of *c*. (*C*) Overall view of the rat Tg gene. The exons are indicated by vertical boxes; introns, by the continuous line; and highly repetitive sequences, by black circles. The broken line represents the segment of the intron whose length has not been determined. The segment of the gene containing the last 17 exons has been described in detail (11).

fragments of these two latter phages were subcloned, and their nucleotide sequence was determined (see below).

The genomic clones listed in Fig. 1 account for at least 70 kb of the rat chromosomal DNA, and they contain almost 6 kb of coding information for the 5′ end of Tg mRNA. By electron microscopic analysis (data not shown), Southern hybridization with specific cDNA clones, and direct DNA sequencing, we determined that 25 exons in these 70 kb show an average size of 200 bp, as previously also found for the 3′ half of the gene (12). The two contiguous exons 9 and 10 are 1100 and 620 bp long, respectively (Table 1). Exons of comparable length have been found in the same position in the human Tg gene (11). We have already demonstrated that the 3′ proximal 3000 bp of the rat Tg mRNA are contained in a genomic segment that is at least 100 kb long divided into 17 exons. Thus, the entire rat Tg gene is organized in 42 exons that are spread out over a chromosomal region that is at least 170 kb long.

Structural Features of the 5′ End of the Rat Tg Gene. To verify that the true 5′ end of the Tg gene is contained in phage λ63 (Fig. 1), we have determined the nucleotide sequence of the restriction fragments hybridizing to rat Tg mRNA. We then compared our sequences with those at the 5′ end of human (6) and bovine (7) mRNA. In our restriction fragments we found a sequence that corresponds to the first 200 nucleotides of both human and bovine mRNA (Fig. 2). By using this sequence homology, we were able to map the location of the first two exons and the translation initiation codon. To prove that the start of transcription for the rat Tg gene is contained within this sequence, we performed the primer extension experiment shown in Fig. 3. This experiment showed that the major start of transcription is about 38–39 nucleotides upstream from the initiator methionine and 29–30 nucleotides downstream from a canonical "TATA" box. Interestingly, a minor start was detected that is 27 nucleotides downstream from a degenerated TATA box (T-G-A-T-A). The sequence of the 5′ end of the gene also shows a structural motif G-G-G-A-C-$\frac{T}{A}$, which is present three times between the TATA box and the first AUG.

A closer inspection of the sequence of the first exon revealed the presence of two contiguous ATG codons at the site of translation initiation, followed by a 16-amino-acid-long hydrophobic segment, as expected for a secretory protein

FIG. 2. Nucleotide sequence of the 5′ terminus of the rat Tg gene. The TATA box is underlined as well as the repeating motif G-G-G-A-C-$\frac{T}{A}$. The arrows indicate the start of the transcription as indicated by the experiment shown in Fig. 3. Coding sequences are in uppercase letters. The boxed ATG represents the putative translation initiation codon, the boxed TAC indicates the codon for the amino-terminal hormonogenic tyrosine. We overscored the sequence complementary to the synthetic oligonucleotide used in the primer elongation experiment shown in Fig. 3. The deduced amino acid sequence is shown below the DNA sequence, followed in order by the amino acid sequence deduced from human and bovine mRNAs. The black circles indicate amino acid sequence homology. We boxed the run of G/T found in the first intron.

(19). Only one methionine is present at the amino terminus of either bovine or human Tg (6, 7). The role of this hydrophobic peptide is supported from earlier experiments showing that Tg has an amino-terminal signal sequence for secretion (20). Furthermore, it has been suggested that the amino-terminal residue of the mature protein from several species is asparagine or aspartic acid (21). An asparagine residue is present, in the three species examined in Fig. 2, at position 20, which may correspond to position 21 in the rat because of the additional methionine residue present at the beginning of the protein.

Among the three species, the putative signal peptide is much less conserved than the mature protein (Fig. 2), as has also been observed in interspecies comparison of other genes (22–24).

In the first intron of the rat Tg gene, we found a sequence 50 nucleotides long made of alternating dG-dT, a sequence that can form Z-DNA (25, 26). Recently it has been shown

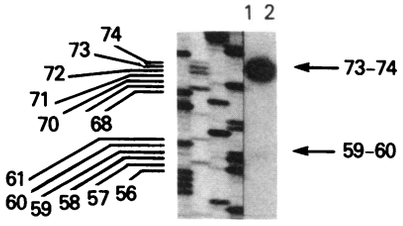

FIG. 3. Primer extension analysis of the 5′ end of the rat Tg gene. A synthetic, kinase-labeled oligonucleotide complementary to nucleotide 47–70 of the rat Tg mRNA (see Fig. 2) was hybridized to 20 μg of either rat thyroid (lane 1) or rat liver (lane 2) total RNA. The hybridization and the extension of the hybridized primer with reverse transcriptase was performed as described (18). The extended primer was electrophoresed on a 6% polyacrylamide/8 M urea sequencing gel. Alongside the extended primer we ran a sequence mix obtained with the chain-terminator method (19) to obtain accurate size markers. An autoradiogram of the dried gel is shown. The major extended product is 73 or 74 nucleotides long. A minor product, 59 or 60 nucleotides long, is also visible.

that such a sequence may stimulate transcription from the simian virus 40 promoter (27). Finally, the amino acid sequence encoded by the second exon reveals the presence of a sequence coding for a peptide (Asn-Ile-Phe-Gly-Tyr-Gln-Val-Asp...) that also occurs in bovine and human Tg and represents a major site for thyroxine formation (4). This peptide is located very close to the amino terminus of the mature protein (6, 7).

Repetitive Units at the 5′ End of the Tg Gene. The size of the exons of the rat Tg gene appears quite homogeneous throughout the entire gene. However, two exceptions are present at the 5′ end of the gene: exons 9 and 10 (see Table 1). The DNA sequence analysis of those two exons revealed the presence of a 4-amino-acid-long peptide (Cys-Trp/Tyr-Cys-Val) present twice in exon 9 and once in exon 10. This peptide represents the core of a longer repeating unit that we have also found in exons 4 and 8 (Fig. 4). This unit, previously shown to occur in the amino-terminal portion of both human and bovine Tg, has already been analyzed in great detail (6, 7). In Fig. 4*B* we show the localization of the repeated elements (black areas) within the 5′-end exons of the rat gene (open boxes). The distribution of the repeated units within the exons suggests that the 5′-end exons of the rat gene arose by serial duplication of an ancestral sequence. The origin of exons or exon segments nonhomologous to the repeated units remains to be assessed.

The presence of two sequences coding for thyroxine-forming sites at the two ends of the rat gene prompted us to search for sequence homology among them. Because a visual inspection of the protein or the DNA sequence of the two sites did not show any obvious homology, we performed a computer search using the NUCALN program (28). We reasoned that if the sequences coding for the functional sites were related, then they should be more homologous to each other than to any other sequence in the rat Tg cDNA.

To verify this hypothesis, we aligned the DNA sequence coding for residues 21–60 of rat Tg, which includes the amino-terminal hormonogenic site (Fig. 2), to the sequence coding for the last 900 amino acids of the protein, which

Table 1. Size of exons and introns in the 5′ half of rat Tg gene

Recombinant λ phages	Exon	Length, bp Exon	Intron
10/63	1	124	450
10/63	2	101	1133 ± 75
10/63	3	112 ± 9	1121 ± 94
10/63	4	124 ± 6	197 ± 20
10/63	5	144 ± 12	9496 ± 1231
10/106	6	127 ± 16	771 ± 51
10/106	7	133 ± 6	184 ± 11
10/106	8	173 ± 5	4479 ± 214
10/106/18	9	1101	751 ± 38
10/106/18	10	628	1352 ± 88
10/106/18	11	244 ± 11	4577 ± 337
18/16	12	116 ± 11	329 ± 44
18/16	13	90 ± 8	538 ± 77
18/16	14	112 ± 17	2948 ± 109
18/16	15	123 ± 12	1100 ± 40
18/16	16	207 ± 23	1532 ± 86
18/16	17	204 ± 26	1438 ± 78
18/16	18	240 ± 23	1532 ± 86
49	19	180 ± 61	1438 ± 78
49/43	20	185 ± 29	>6000
49/43	21	171 ± 24	2426 ± 388
49/43	22	210 ± 74	~5000
2	23	193 ± 37	6399 ± 1425
2	24	203 ± 18	5816 ± 1100
2	25	176 ± 27	

The size of exons and introns have been estimated by electron microscopy. Exons 1, 2, 9, and 10 have also been sized by direct DNA sequencing. The size of intron 22 has been deduced by blot hybridization. The phage(s) used in each case for R-looping are indicated.

contains the carboxyl-terminal site (8). In five attempts the sequence coding for the amino-terminal site was aligned to sequences different from the one coding for the carboxyl-terminal site. In each attempt we removed from the long carboxyl-terminal sequence the segment that had shown homology in the previous alignment. In each case the homologies involved were not >35% (data not shown).

DISCUSSION

The rat gene for Tg is, together with the gene for factor VIII (29), the largest eukaryotic gene isolated so far. The amazingly complex exon–intron structure of the Tg gene seems to confirm the direct relationship found between total size of exons and total size of introns found in a large number of eukaryotic genes (30). Also the homogeneous size of the exons is not unusual in that most of the exons of eukaryotic genes show a size between 50 and 200 bp (30).

As in most genes, the introns show a wide size range. In an attempt to gain some insight into the evolutionary history of the Tg gene, we concentrated our attention on the sequences coding for the functional sites, i.e., the sites where a diiodinated tyrosyl residue couples with another diiodinated tyrosyl residue to form the thyroid hormone, thyroxine. Despite the large size of Tg, only two molecules of hormone are formed per subunit (2). Very specific sites in Tg seem to be involved in hormone synthesis (4, 5); it has been suggested that the specificity, in each coupling pair, is restricted to only one of the two coupling residues (8).

Two major polypeptide sites for thyroxine synthesis have been isolated from Tg, sequenced (4, 5), and subsequently mapped with the help of the sequence deduced from the mRNA sequence (6–9). They are localized at the two extremities of the protein. In the rat gene, the two coding sequences are localized in exons 2 (site 1) and 38 (site 2), respectively. Surprisingly, no homology can be demonstrated between these two highly specialized sites, either at the protein or at the DNA sequence level.

We previously localized in the rat Tg mRNA sequence, very close to site 2, a third low-affinity thyroxine-forming site (5). The comparison between sites 2 and 3 shows a different protein sequence, but the two sites share a common DNA sequence (12). A weak homology to this sequence is also present once per exon in the last seven exons of the rat gene (12). We then proposed that the 3′-terminal region of the Tg gene arose by duplication of an ancestral sequence related to the sequence coding for sites 2 and 3. We report in this paper that the 5′ end of the gene is also made of repeating units, but these are unrelated to the thyroxine-forming site present in the 5′ region. We could find no evidence for the presence of this repeating unit at the 3′ end of the gene. Thus, the two extremities of the rat Tg gene display common features

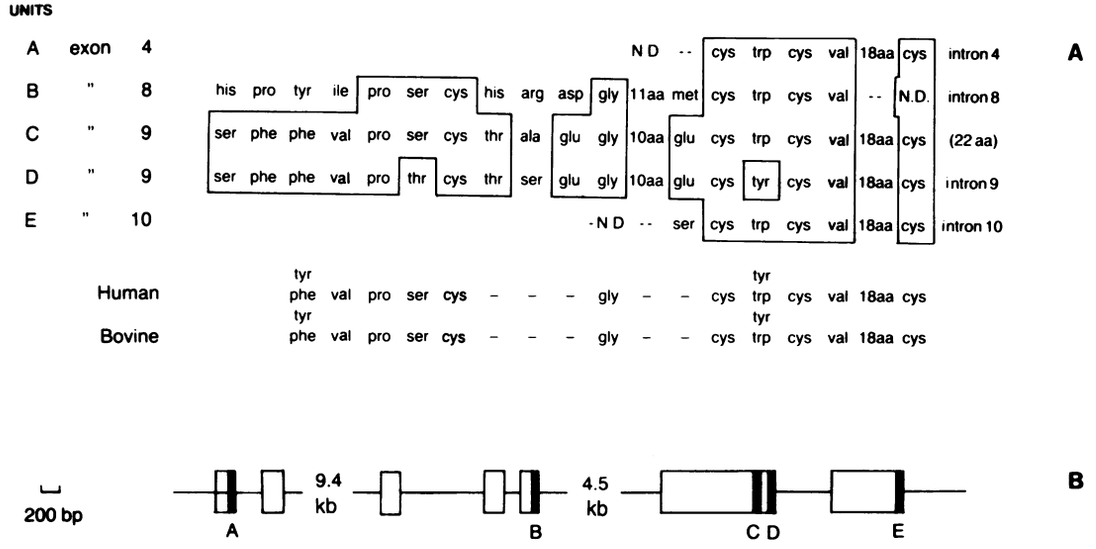

FIG. 4. Alignment of repeating domains present in the amino-terminal region of Tg. (*A*) Identical residues between repeating units A, B, C, D, and E are boxed. Consensus sequences derived from the sequence of human and bovine Tg are shown below. (*B*) Localization of the repeating unit in the exons of the rat gene. The black areas represent the repeating domains, the open boxes are exons, and the continuous lines are introns. ND, sequence not determined.

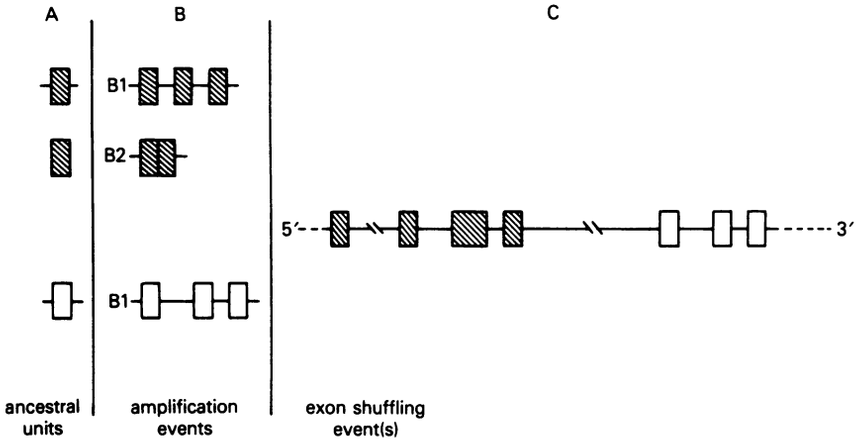

FIG. 5. A model for the origin of the Tg gene. The hatched boxes represent the repeated units at the 5' end of the gene. The empty boxes are the units at the 3' end. (*A*) Ancestral unit before duplication. (*B*) Ancestral units after duplication, with (B1) or without (B2) duplication of flanking sequences. (*C*) Fusion of genomic segments containing duplicated units.

(presence of repeating units, presence of a thyroxine-forming site) but do not show any sequence homology. We suggest that both regions originated by serial duplication of a different ancestral sequence and that they evolved independently toward a common function—i.e., the ability to code for a thyroxine-forming site (Fig. 5).

In conclusion, what seems to be unique in the structure of the Tg gene is that different segments of the gene encode similar specific functions but share no sequence homology. It has been proposed that eukaryotic genes may evolve rapidly by shuffling of exons coding for preformed protein domains that could be utilized by different genes (31). The most evident example of such a mechanism is the recently elucidated structure of the gene for the low density lipoprotein (LDL) receptor (32), whose exons share sequence homology with several different genes. The structure of the Tg gene may be interpreted instead as the result of two independent evolutionary attempts to build a thyroxine-forming site. The two sites may have been fused subsequently on the same polypeptide chain. Analysis of the structure of the thyroglobulin gene in lower species may support or falsify our hypothesis. We would predict that it should be possible to find an organism in which the two (functional?) thyroxine-forming sites are on two separate polypeptide chains.

We thank Maxine Singer, Bruce Paterson, and Tom Fanning for reviewing the manuscript and Jacek Skowronski for helpful suggestions. We also thank Rita Cerillo for excellent technical assistance and Gail Gray for editing the manuscript. This work was supported in part by Progetto Finalizzato Ingegneria Genetica e basi molecolari delle Malattie Ereditarie and by Progetto Finalizzato Oncologia del Consiglio Nazionale delle Ricerche.

1. Edelhoc, H. & Robbins, J. (1978) in *The Thyroid*, eds. Werner, S. C. & Ingbar, S. H. (Harper & Row, New York) 4th Ed., pp. 62–76.
2. Salvatore, G. & Edelhoc, H. (1973) in *Hormonal Proteins and Peptides*, ed. Choh, H. L. (Academic, New York), pp. 201–240.
3. Lissitzky, S. (1984) *J. Endocrinol. Invest.* **7**, 65–76.
4. Rawitch, A. B., Chernoff, S. B., Litwer, M. R., Rouse, J. B. & Hamilton, J. W. (1983) *J. Biol. Chem.* **285**, 2079–2082.
5. Marriq, C., Rolland, M. & Lissitzky, S. (1982) *EMBO J.* **1**, 397–401.
6. Malthiery, Y. & Lissitzky, S. (1985) *Eur. J. Biochem.* **147**, 53–58.
7. Mercken, L., Simons, M. J., De Martinoff, G., Swillens, S. & Vassart, G. (1985) *Eur. J. Biochem.* **147**, 59–64.
8. Di Lauro, R., Obici, S., Condliffe, S., Ursini, M. V., Musti, A. M., Moscatelli, C. & Avvedimento, V. E. (1985) *Eur. J. Biochem.* **148**, 7–11.
9. Mercken, L., Massaer, M., Simons, M. J., Swillens, S. & Vassart, G. (1984) *Biochem. Biophys. Res. Commun.* **125**, 961–966.
10. Van Ommen, G. J. B., Arneberg, A. C., Baas, F., Brocas, H., Sterk, A., Tegelaers, W. H., Vassart, G. & de Vijlder, J. J. (1983) *Nucleic Acids Res.* **11**, 2273–2275.
11. Targnovik, V. M., Pohl, V., Cristophe, D., Cabrer, B., Brocas, H. & Vassart, G. (1984) *Eur. J. Biochem.* **141**, 271–277.
12. Avvedimento, V. E., Musti, A. M., Obici, S., Cocozza, S. & Di Lauro, R. (1984) *Nucleic Acids Res.* **12**, 3461–3472.
13. Cristophe, D., Pohl, V., Van Heuverswijn, B., De Martinoff, G., Dumont, J. E., Pasteels, L. J. & Vassart, G. (1982) *Biochem. Biophys. Res. Commun.* **105**, 1166–1175.
14. Di Lauro, R., Avvedimento, V. E., Cerillo, R., Cocozza, S., Condliffe, D., Monticelli, A., Musti, A. M., Obici, S., Ursini, V. & Varrone, S. (1985) in *Thyroglobulin: The Prothyroid Hormone*, eds. Eggo, M. C. & Burrow, G. N. (Raven, New York), pp. 77–85.
15. Sargent, T. D., Wu, J. R., Sala-Trepat, J. M., Wallace, R. B., Reyes, A. A. & Bonner, J. (1979) *Proc. Natl. Acad. Sci. USA* **76**, 3256–3260.
16. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Laboratory, Cold Spring Harbor, NY).
17. Davis, R. W., Simon, M. & Davidson, N. (1971) *Methods Enzymol.* **21D**, 413–428.
18. Sanger, F., Nicklen, S. & Coulson, A. R. (1977) *Proc. Natl. Acad. Sci. USA* **74**, 5463–5467.
19. Blobel, G., Walter, P., Chang, C. N., Goldman, B. M., Erikson, A. M. & Lingappa, V. R. (1979) in *Secretory Mechanisms*, eds. Hopkins, C. R. & Duncan, C. J. (Cambridge Univ. Press, Cambridge), pp. 9–36.
20. Alvino, C. G., Tassi, V., Polistina, C., Di Lauro, R. & Bonatti, S. (1982) *Eur. J. Biochem.* **125**, 15–19.
21. Spiro, M. J. (1970) *J. Biol. Chem.* **245**, 5820–5826.
22. Hobart, P., Crawford, R., Shen, L. P., Pictet, R. & Rutter, W. J. (1980) *Nature (London)* **288**, 137–141.
23. Shen, L. P. & Rutter, W. J. (1984) *Science* **224**, 168–171.
24. Uhler, M. & Herbert, E. (1983) *J. Biol. Chem.* **258**, 257–261.
25. Haniford, D. B. & Pulleybank, D. E. (1983) *Nature (London)* **302**, 632–634.
26. Nordheim, A. & Rich, A. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 1821–1825.
27. Hamada, H., Seidman, M., Howard, B. H. & Gorman, C. (1984) *Mol. Cell. Biol.* **4**, 2622–2630.
28. Wilbur, W. J. & Lipman, D. J. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 726–730.
29. Gitschier, J., Wood, W. I., Goralka, T. M., Wion, K. L., Chen, E. Y., Eaton, D. H., Vehar, G. A., Capon, D. J. & Lawn, R. M. (1984) *Nature (London)* **312**, 326–330.
30. Naora, H. & Deacon, N. J. (1982) *Proc. Natl. Acad. Sci. USA* **79**, 6196–6200.
31. Gilbert, W. (1985) *Science* **228**, 823–824.
32. Shudhof, T. C., Goldstein, J. L., Brown, M. S. & Russell, D. W. (1985) *Science* **228**, 815–822.