

Primary structure of human thyroglobulin deduced from the sequence of its 8448-base complementary DNA

Yves MALTHIÉRY and Serge LISSITZKY†

Laboratoire de Biochimie Médicale associé au Centre National de la Recherche Scientifique (Unité Associée 178) et Unité 38 de l'Institut National de la Santé et de la Recherche Médicale, Faculté de Médecine, Marseille

(Received December 17, 1986/February 16, 1987) — EJB 86 1361

The mRNA encoding human thyroglobulin has been cloned and sequenced. It is made up of a 8301-nucleotide segment encoding a preprotein monomer of 2767 amino acids, flanked by non-coding 5' and 3' regions of 41 and 106 nucleotides, respectively. This preprotein consists of a leader sequence of 19 amino acids, followed by the sequence of the mature monomer, corresponding to a polypeptide of 2748 amino acids ($M_r = 302\,773$). On its amino-terminal side, 70% of the monomer is characterized by the presence of three types of repetitive units. In contrast, the remaining 30% of the protein is devoid of repetitive units. This last region however shows an interesting homology (up to 64%) with the acetylcholinesterase of *Torpedo californica*. The sites of thyroid hormones synthesis are clustered at both ends of the thyroglobulin monomer. By contrast, the potential glycosylation sites are scattered along the polypeptide chain.

Thyroglobulin is a protein specifically synthesised by the thyroid gland, and constitutes the support for the production of the two thyroid hormones, thyroxine and triiodothyronine [1]. The existence of thyroglobulin was demonstrated a century ago [2], but its structure has been elucidated only recently. It is a dimeric glycoprotein with an M_r of 660 000, of two identical subunits [3, 4] encoded by a single mRNA with a sedimentation coefficient of 33 S (8500 nucleotides) [5–7]. Thyroglobulin is synthesised by the thyrocyte, then exported to the vesicular lumen where its maturation begins by the iodination of several tyrosine residues, and coupling of some of the iodotyrosine residues [8]. Then, by an endocytotic process, the molecule is absorbed into the thyrocyte where several selective cleavages occur in the lysosomes, resulting in the release of the thyroid hormones, and complete degradation of the rest of the molecule.

For a thyroglobulin iodine content of 0.5% (which is rarely attained in man) a maximum of 3.5 hormonal residues per thyroglobulin molecule are formed through a reaction catalyzed by the enzyme thyroid peroxidase [9]. Four hormone-synthesis sites have been described, corresponding to four tyrosine residues in fixed positions [10–12].

The structure of human thyroglobulin seems to be responsible for the specific fixation of iodine, and the production of functional thyroid hormones. Several human pathologies are associated with an abnormal thyroid function. Since the recent demonstration of the implication of a defect in thyroglobulin gene structure in the development of congenital goitre in cattle

[13], it is likely that knowledge of the structure of human thyroglobulin mRNA will help to elucidate the structural bases of human thyroid pathologies. We describe here the complete nucleotide sequence of human thyroglobulin mRNA.

MATERIALS AND METHODS

Preparation and sequencing of DNA

cDNA fragments corresponding to human thyroglobulin mRNA were prepared from recombinant plasmids named M1–M4 and B2–B4 (see Fig. 1), as previously described [7, 14]. Two additional clones, named B1 (kind gift of H. Brocas and G. Vassart) and M5, were constructed by G-C tailing of cDNA, and eventual insertion into the *Pst*I site of pBR322, according to Maniatis et al. [15]. Restriction endonucleases were used as recommended by the suppliers. Fragments carrying 5'-protruding ends were labeled using alkaline phosphatase (CIP Boehringer) and T4 polynucleotide kinase (BRL) with [^{32} P]ATP (3000 Ci/mmol, Amersham). Fragments carrying 3'-protruding ends were labeled with cordycepin (3'-deoxyadenosine) or 2',3'-dideoxyadenosine 5'-[^{32} P]-phosphate ([^{32} P]ddATP, 3000 Ci/mmol, Amersham) in the presence of terminal deoxynucleotidyl transferase (BRL).

The labeled fragments were isolated and sequenced according to the method described by Maxam and Gilbert [16].

mRNA preparation

Human thyroglobulin mRNA was extracted by the guanidine · HCl procedure [17] or by the guanidinium thiocyanate/CsCl gradient procedure [18], from a Graves' disease thyroid obtained surgically. A single passage through oligo(dT)-cellulose was used to prepare the fraction enriched in poly(A)-containing RNA. The quality of the RNA preparation was monitored by electrophoresis on agarose/methylmercury-hydroxide gels.

Correspondence to Y. Malthiéry, Laboratoire de Biochimie Médicale, Faculté de Médecine, 27 Boulevard Jean Moulin, F-13385 Marseille Cedex 5, France

† This paper is dedicated to the memory of Professor Serge Lissitzky.

Enzymes. Reverse transcriptase or RNA-directed DNA nucleotidyltransferase (EC 2.7.7.49); terminal deoxynucleotidyltransferase (EC 2.7.7.31); T4 polynucleotide kinase (EC 2.7.1.78); alkaline phosphatase (EC 3.1.3.1); restriction endonucleases *Bgl*II, *Eco*RI, *Pst*I and *Sau*3A1 (EC 3.1.21.4).

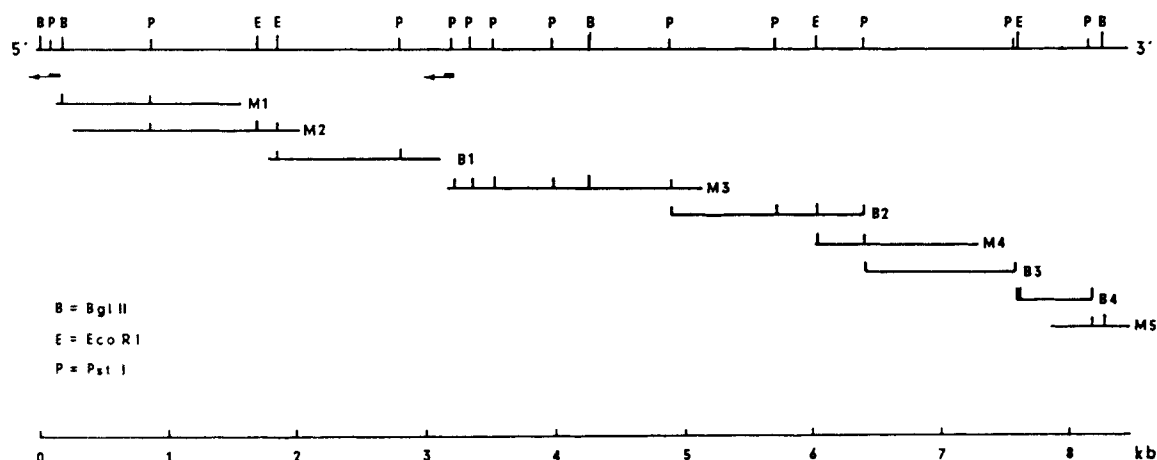


Fig. 1. Restriction map of DNA coding for human thyroglobulin obtained from nine scored cDNA clones. The sequencing of the 5' end and the region between clones B1 and M3 were obtained by primer-extension. The primers were isolated from clones M1 and M3, respectively

Primer-extension experiments

The 5'-end extension of clone M1 has been already described [19]. A gap between clones M3 and B1 was filled up by primer-extension: a 147-bp restriction fragment corresponding to position 35–182 of clone M3 was isolated by polyacrylamide gel electrophoresis, and subsequently cut by *Sau3A1* (position 111 of M3). The fragment was labelled with ^{32}P at the 5' end as described above, and the two complementary strands separated on a 8% acrylamide/urea gel, and eluted according to Maxam and Gilbert [16].

Poly(A)-rich mRNA (50 μg) was hybridized with 50 ng single-stranded primer DNA labeled at its 5'-end (1:1 ratio of homologous sequences) in 100 μl 0.4 M NaCl, 10 mM Pipes pH 6.4 for 4 h at 65°C. Ethanol precipitation was performed following hybridization. Extension was carried out for 45 min at 42°C in 0.1 M Tris/HCl pH 8.3, 0.14 M KCl, 10 mM MgCl_2 , 28 mM 2-mercaptoethanol in the presence of 1 mM dATP, dCTP, dGTP, dTTP and 40 units of reverse transcriptase (PL) in a final volume of 50 μl . The mRNA was then hydrolysed by addition of 25 μl 0.15 M NaOH for 60 min at 65°C. The reaction was stopped by neutralization and the ss-cDNA was precipitated by cold ethanol. Sequencing of the ss-cDNA was performed according to Maxam and Gilbert [16].

Sequence analysis

Results were analyzed using several computer programs developed by B. Jacq and B. Belon (unpublished) and B. Malthiéry et al. [20]. These programs allowed the restriction analysis of nucleic acid sequences, the search for homologies in nucleotide and protein sequences and the study of hydrophilicity.

RESULTS

The complete nucleotide sequence of the human thyroglobulin mRNA was deduced from the sequence of nine overlapping recombinant plasmids, corresponding to 98% of the mRNA sequence, and for the remaining 2% by sequencing primer extensions, corresponding to a gap between two clones, and to the 5'-end of the message (Fig. 1). The nine clones

were inserted into the plasmid vector pBR322, by G-C tailing (clones M1–4 [7] and B1) or by ligation of cohesive ends (clones B2–4 [14]). Restriction mapping of the inserts, and cross-hybridization experiments allowed the respective positioning of the inserts. The sequences were established unambiguously by sequencing both DNA strands. Two regions could not be cloned: a 56-nucleotide fragment between clones B1 and M3, and a 156-nucleotide fragment at the 5'-end of the mRNA. Those regions were covered by primer extension, as described in Methods. The sequence of the latter was verified by sequencing an appropriate genomic clone, and by comparison to the sequence published by others [21]. The sequence around the junction between clones B3 and B4 was monitored by sequencing a genomic subclone covering that region (generous gift of F. Baas).

The mRNA encoding human thyroglobulin is a polynucleotide chain of 8448 nucleotides, provided with a poly(A) tail. It consists of a 5' non-translated segment of 41 nucleotides, followed by a coding sequence of 8301 nucleotides with a single open reading frame, and a 3' non-translated segment of 106 nucleotides, showing the canonical polyadenylation signal AAUAAA in position 8427 (Fig. 2). A polypeptide sequence of 2767 amino acids can be derived from the nucleotide sequence. A leader peptide of 19 amino acids is followed by a 2748-amino-acid polypeptide, corresponding to the monomeric human thyroglobulin. The deduced M_r of that polypeptide (302773) agrees with the size of the polypeptide obtained by *in vitro* translation of human thyroglobulin mRNA [22] and with the size of the glycosylated monomeric thyroglobulin [23].

The amino-acid composition is similar to that of the mature protein [24], and confirms a rather high amount of serine and glycine residues (9.7% and 7.6% respectively), and a small proportion of lysine (3.3%; Table 1). Hydrophobic and charged amino-acid residues are homogeneously distributed on the polypeptide, whereas the cysteines involved in the repetitive structures (see below) and for the tyrosines are not. Each monomeric chain contains 67 tyrosine residues, representing 2.44% of total. They are mostly clustered in the repetitive structures and in a region near the C-terminal, where they constitute up to 9% of total amino acids.

20 potential *N*-glycosylation sites (Asn-Xaa-Thr/Ser) are present in the protomeric sequence, that is twice the number

1 ATG-GCC-CTG-GTC-CTG-GAG-ATC-TTC-ACC-CTG-CTG-GCC-TCC-ATC-TGC-TGG-GTG-TCG-GCC-AAT-ATC-TTC-GAG-TAC-CAG-GTT-GAT-GCC-CAG-CCC-
 -19 aet-ala-leu-val-leu-glu-ile-phe-thr-leu-leu-ala-ser-ile-cys-trp-val-ser-ala-asn-ile-phe-gly-tyr-gln-val-asp-ala-gln-pro-
 91 CTT-CGT-CCC-TGT-GAG-CTG-CAG-AGG-GAA-ACC-GCC-TTT-CTG-ARG-CAR-GCA-GAC-TAC-GTG-CCC-CAG-TGT-GCA-GAG-GAT-GGC-AGC-TTC-CAG-ACT-
 12 leu-arg-pro-cys-glu-leu-gln-arg-glu-thr-ala-phe-leu-lys-gln-ala-asp-tyr-val-pro-gln-cys-ala-glu-asp-gly-ser-phe-gln-thr-
 181 GTC-CAG-TGC-CAG-ACC-GAG-GGC-TCC-TGC-TGG-TGT-GTG-GGT-GCC-ACC-GGC-AGT-GAA-GTG-CTG-GGC-AGC-AGG-CAG-CCA-GGA-CGG-CCT-GTG-
 42 val-gln-cys-gln-asn-asp-gly-arg-ser-cys-trp-cys-val-gly-ala-asn-gly-ser-glu-val-leu-gly-ser-arg-gln-pro-gly-arg-pro-val-
 271 GCT-TGT-CTG-TCA-TTT-TGT-CAG-CTA-CAG-AAA-CAG-CAG-ATC-TTA-CTG-AGT-GGC-TAC-ATT-ACC-AGC-ACA-GAC-ACC-TCC-TAC-CTC-CCT-CAG-TGT-
 72 ala-cys-leu-ser-phe-cys-gln-leu-gln-lys-gln-gln-ile-leu-leu-ser-gly-tyr-ile-asn-ser-thr-asp-thr-ser-tyr-leu-pro-gln-cys-
 361 CAG-GAT-TCA-GGG-GAC-TAC-GCG-CCT-GTT-CAG-TGT-GAT-GTG-CAG-CAT-GTC-CAG-TGC-TGG-TGT-GTG-GAC-GCA-GAG-GGG-ATG-GAG-GTG-TAT-GGG-
 102 gln-asp-ser-gly-asp-tyr-ala-pro-val-gln-cys-asp-val-gln-his-val-gln-cys-trp-cys-val-asp-ala-glu-gly-met-glu-val-tyr-gly-
 451 ACC-CGC-CAG-CTG-GGG-AGG-CCA-AGG-CGA-TGT-CCA-AGG-AGC-TGT-GAA-ATA-AGA-AAT-CGT-CGT-CTT-CTC-CAC-GGG-GTG-GGA-GAT-AGT-TCA-CCA-
 132 thr-arg-gln-leu-gly-arg-pro-lys-arg-cys-pro-arg-ser-cys-glu-ile-arg-asn-arg-arg-leu-leu-his-gly-val-gln-asp-lys-ser-pro-
 541 CCC-CAG-TGT-TCT-GCG-GAG-GAG-TTT-ATG-CCT-GTC-CAG-TGC-AAA-TTT-GTC-ACC-ACC-ACA-GAC-ATG-ATG-ATT-TTT-GAT-CTG-GTC-CAC-ACC-
 162 pro-gln-cys-ser-ala-glu-gly-glu-phe-met-pro-val-gln-cys-lys-phe-val-asn-thr-thr-asp-met-met-ile-phe-asp-leu-val-his-ser-
 631 TAC-ACC-AGG-TTT-CCA-GAT-GCA-TTT-GTC-AGC-TTC-AGT-TCC-TTC-CAG-AGG-AGG-TTC-CCT-GAG-GTA-TCT-GGG-TAT-TGC-CAC-TGT-GCT-GAC-AGC-
 192 tyr-asn-arg-phe-pro-asp-ala-phe-val-thr-phe-ser-ser-phe-gln-arg-arg-phe-pro-glu-val-ser-gly-tyr-cys-his-cys-ala-asp-ser-
 721 CAA-GGG-CGG-GAA-CTG-GCT-GAG-ACA-GGT-TTG-GAG-TTG-TTA-CTG-GAT-GAA-ATT-TAT-GAC-ACC-ATT-TTT-GCT-GGC-CTG-GAC-CTT-CCT-TCC-ACC-
 222 gln-gly-arg-glu-leu-ala-glu-thr-gly-leu-glu-leu-leu-leu-asp-glu-ile-tyr-asp-thr-ile-phe-ala-gly-leu-asp-leu-pro-ser-thr-
 811 TTC-ACT-GAA-ACC-ACC-CTG-TAC-CGG-ATA-CTG-CAG-AGA-CGG-TTC-CTC-GCA-GTT-CAR-TAC-GTC-ATC-TCT-GGC-AGA-TTC-CGA-TGC-CCC-ACA-AAA-
 252 phe-thr-glu-thr-thr-leu-tyr-arg-ile-leu-gln-arg-arg-phe-leu-ala-val-gln-ser-val-ile-ser-gly-arg-phe-arg-cys-pro-thr-lys-
 901 TGT-GAA-GTG-GAG-CGG-TTT-ACA-GCA-ACC-AGC-TTT-GGT-CAC-CCC-TAT-GTT-CCA-AGC-TGC-CGC-CGA-AAT-GGC-GAC-TAT-CAG-GCG-GTG-CAG-TGC-
 282 cys-glu-val-glu-arg-phe-thr-ala-thr-ser-phe-gly-his-pro-tyr-val-pro-ser-cys-arg-arg-asn-gly-asp-tyr-gln-ala-val-gln-cys-
 991 CAG-ACG-GAA-GGG-CCC-TGC-TGG-TGT-GTG-GAC-GCC-CAG-GGG-ARG-GAA-ATG-CAT-GGA-ACC-CGG-CAG-CAA-GGG-CAG-CCG-CCA-TCT-TGT-GCT-GAA-
 312 gln-thr-glu-gly-pro-cys-trp-cys-val-asp-ala-gln-gly-lys-glu-met-his-gly-thr-arg-gln-gln-gly-glu-pro-pro-ser-cys-ala-glu-
 1081 GGC-CAA-TCT-TGT-GCC-TCC-GAA-AGG-CAG-CAG-GCC-TTG-TCC-AGA-CTC-TAC-TTT-GGG-ACC-TCA-GGC-TAC-TTC-AGC-CAG-CAC-GAC-CTG-TTC-TCT-
 342 gly-gln-ser-cys-ala-ser-glu-arg-gln-gln-ala-leu-ser-arg-leu-tyr-phe-gly-thr-ser-gly-tyr-phe-ser-gln-his-asp-leu-phe-ser-
 1171 TCC-CCA-GAG-AAA-AGA-TGG-GCC-TCT-CCA-AGA-GTA-GCC-AGA-TTT-GCC-ACA-TCC-TGC-CCA-CCC-ACG-ATC-AGG-GAG-CTC-TTT-GTG-GAC-TCT-GGG-
 372 ser-pro-glu-lys-arg-trp-ala-ser-pro-arg-val-ala-arg-phe-ala-thr-ser-cys-pro-pro-thr-ile-lys-glu-leu-phe-val-asp-ser-gly-
 1261 CTT-CTC-CGC-CCA-ATG-GTG-GAG-GGA-CAG-AGC-CAA-CAG-TTT-TCT-GTC-TCA-GAA-AAT-CTT-CTC-AAA-GAA-GCC-ATC-CCA-GCA-ATT-TTT-CCC-TCC-
 402 leu-leu-arg-pro-met-val-glu-gly-gln-ser-gln-gln-phe-ser-val-ser-glu-asn-leu-leu-lys-glu-ala-ile-arg-ala-leu-phe-pro-ser-
 1351 CGA-GGG-CTG-GCT-CGT-CTT-GCC-CTT-CAG-TTT-ACC-ACC-ACC-CCA-AGG-AGA-CTC-CAG-CAA-ACC-CIT-TTT-GGA-GGG-AAA-TTT-TTG-GTG-AAT-GTT-
 432 arg-gly-leu-ala-arg-leu-ala-leu-gln-phe-thr-thr-asn-pro-lys-arg-leu-gln-gln-asn-leu-phe-gly-gly-lys-phe-leu-val-asn-val-
 1441 GGC-CAG-TTT-ACC-TTG-TCT-GGA-GCC-CTT-GGC-ACA-AGA-GGC-ACA-TTT-ACC-TTC-AGT-CAR-TTT-TTC-CAG-CAA-CTT-GGT-CTT-GCA-AGC-TTC-TTG-
 462 gly-gln-phe-asn-leu-ser-gly-ala-leu-gly-thr-arg-gly-thr-phe-asn-phe-ser-gln-phe-phe-gln-gln-leu-gly-leu-ala-ser-phe-leu-
 1531 AAT-GGA-GGG-AGA-CAA-GAA-GAT-TTG-GCC-AGG-CCA-CTC-TCT-GTG-GGA-TTA-GAT-TCA-AAT-TCT-TCC-ACA-GGA-ACC-CCT-GAA-GCT-GCT-AGG-AGG-
 492 asn-gly-gly-arg-gln-glu-asp-leu-ala-lys-pro-leu-ser-val-gly-leu-asp-ser-asn-ser-ser-thr-gly-thr-pro-glu-ala-ala-lys-lys-
 1621 GAT-GGT-ACT-ATG-AAT-ARG-CCA-ACT-GTG-GGC-AGC-TTT-GGC-TTT-GAA-ATT-ACC-TCA-CAG-GAG-ACC-CAA-AAT-GCC-CTC-AAA-TTC-CTT-GCT-TCT-
 522 asp-gly-thr-met-asn-lys-pro-thr-val-gly-ser-phe-gly-phe-glu-ile-asn-leu-gln-glu-asn-gln-asn-ala-leu-lys-phe-leu-ala-ser-
 1711 CTC-CTG-GAG-CTT-CCA-GAA-TTC-CTT-CTC-TTC-TTG-CAA-CAT-GCT-ATC-TCT-GTG-CCA-GAA-GAT-GTG-GCA-AGA-GAT-TTA-GGT-GAT-GTG-ATG-GAA-
 552 leu-leu-gln-asp-val-pro-leu-phe-leu-phe-leu-gln-his-ala-ile-ser-val-pro-glu-asp-val-ala-arg-asp-leu-gly-ala-val-met-gln-
 1801 ACG-GTA-CTC-GAC-TCC-CAG-ACC-TGT-GAG-CAG-ACA-CCT-GAA-AGG-CTA-TTT-GTC-CCA-TCA-TGC-ACG-ACA-GAA-GGA-AGC-TAT-GAG-GAT-GTC-CAA-
 582 thr-val-leu-asp-ser-gln-thr-cys-glu-gln-thr-pro-glu-arg-leu-phe-val-pro-ser-cys-thr-thr-glu-gly-ser-tyr-glu-asp-val-gln-
 1891 TGC-TTT-TCC-GGA-GAG-TGC-TGG-TGT-GTG-AAT-TCC-TGG-GGC-AAA-CAG-CTT-CCA-GGC-TCA-AGA-GTC-AGA-GAT-GGA-CAG-CCA-AGG-TGC-CCC-ACA-
 612 cys-phe-ser-gly-glu-cys-trp-cys-val-asn-ser-trp-gly-lys-glu-leu-pro-gly-ser-arg-val-arg-asp-gly-gln-pro-arg-cys-pro-thr-
 1981 GAC-TGT-GAA-ARG-CAA-AGG-GCT-GCG-ATG-CAA-AGC-CTC-ATG-GGC-AGC-CAG-CCT-GCT-GGC-TCC-ACC-TTG-TTT-GTC-CCT-GCT-TGT-ACT-AGT-GAG-
 642 asp-cys-glu-lys-gln-arg-ala-arg-met-gln-ser-leu-met-gly-ser-gln-pro-ala-gly-ser-thr-leu-phe-val-pro-ala-cys-thr-ser-glu-
 2071 GGA-CAT-TTC-CTG-CCT-GTC-CAG-TGC-TTC-ACC-TCA-GAG-TGC-TAC-TGT-GTT-GAT-GCT-GAG-GGT-CAG-GCC-ATT-CCT-GGA-ACT-CGA-AGT-GCA-ATA-
 672 gly-his-phe-leu-pro-val-gln-cys-phe-asn-ser-gly-cys-tyr-cys-val-asp-ala-glu-gly-gln-ala-ile-pro-gly-thr-arg-ser-ala-ile-
 2161 GGG-ARG-CCC-ARG-AAA-TGC-CCC-AGC-CCC-TGT-CAA-TTA-CAG-TCT-GAG-CAA-GCT-TTC-CTC-AGG-ACG-GTG-CAG-GCC-CTG-CTC-TCT-ACC-TCC-AGC-
 702 gly-lys-pro-lys-lys-cys-pro-thr-pro-cys-gln-leu-gln-ser-glu-gln-ala-phe-leu-arg-thr-val-gln-ala-leu-leu-ser-asn-ser-ser-
 2251 ATG-CTA-CCC-ACC-CTT-TCC-GAC-ACC-TAC-ATC-CCA-CAG-TGC-AGC-ACC-GAT-GGG-CAG-TGG-AGA-CAA-GTG-CAA-TGC-AAT-GGG-CCT-CCT-GAG-CAG-
 732 met-leu-pro-thr-leu-ser-asp-thr-tyr-ile-pro-gln-cys-ser-thr-asp-gly-gln-trp-arg-gln-val-gln-cys-asn-gly-pro-pro-glu-gln-
 2341 GTC-TTC-GAG-TTG-TAC-CAA-CGA-TGG-GAG-GCT-ERG-ACC-ARG-GGC-CAG-GAT-CTG-ACG-CCT-GCC-ARG-CTG-CTA-GTG-ARG-ATC-ATG-AGC-TAC-AGA-
 762 val-phe-glu-leu-tyr-gln-arg-trp-glu-ala-gln-asn-lys-gly-gln-asp-leu-thr-pro-ala-lys-leu-leu-val-lys-ile-met-ser-tyr-arg-
 2431 GAA-GCA-GCT-TCC-GCA-ACC-TTC-AGT-CTC-TTT-ATT-CAA-AGT-CTG-TAT-GAG-GCT-GGC-CAG-CAA-GAT-GTC-TTC-CCG-GTG-CTG-TCA-CAA-TAC-CCT-
 792 glu-ala-ala-ser-gly-asn-phe-ser-leu-phe-ile-gln-ser-leu-tyr-glu-ala-gly-gln-gln-asp-val-phe-pro-val-leu-ser-gln-tyr-pro-
 2521 TCT-CTG-CAG-GAT-GTC-CCA-CTA-GCA-CTG-CTG-GAA-GGG-AAA-CGG-CCC-CAG-CCC-AGG-GAG-AAT-ATC-CTC-CTG-GAG-CCC-TAC-CTC-TTC-TGG-CAG-
 822 ser-leu-gln-asp-val-pro-leu-ala-ala-leu-glu-gly-lys-arg-pro-gln-pro-arg-glu-asn-ile-leu-leu-glu-pro-tyr-leu-phe-leu-trp-gln-
 2611 ATC-TTA-AAT-GGC-CAA-CTC-AGC-CAA-TAC-CCG-GGG-TCC-TAC-TCA-GAC-TTC-AGC-ACT-CCT-TTG-GCA-CAT-TTT-GAT-CTT-CGG-ARC-TGC-TGG-TGT-
 852 ile-leu-asn-gly-gln-leu-ser-gln-tyr-pro-gly-ser-tyr-ser-asp-phe-ser-thr-pro-leu-ala-his-phe-asp-leu-arg-asn-cys-trp-cys-
 2701 GTG-GAT-GAG-GCT-GGC-CAA-GAA-CTG-GAA-GGA-ATG-CGG-TCT-GAG-CCA-AGC-ARG-CTC-CCA-ACG-TGT-CCT-GGC-TCC-TGT-GAG-GAA-GCA-ARG-CTC-
 882 val-asp-glu-ala-gly-gln-glu-leu-glu-gly-met-arg-ser-glu-pro-ser-lys-leu-pro-thr-cys-pro-gly-ser-cys-glu-glu-ala-lys-leu-

Fig. 2. Nucleotide sequence of human thyroglobulin cDNA, and the deduced sequence of the protein. Numbering starts at nucleotide 42, corresponding to the initiation codon ATG, and at the 20th amino acid residue (Asn), which is the N-terminal residue of the mature protein (asterisk). The first 19 residues constitute the leader sequence of thyroglobulin. The AATAAA polyadenylation signal is underscored. The putative N-glycosylation sites (Asn-Xaa-Ser/Thr) are underscored. The tyrosine residues involved in hormone synthesis are boxed

2791 CGT-GTA-CTG-CAG-TTC-ATT-AGG-GAA-ACG-GAA-GAG-ATT-GTT-TCA-GCT-TCC-AAC-AGT-TCT-CGG-TTC-CCT-CTG-GGG-GAG-AGT-TTC-CTG-GTG-GCC-
 912 arg-val-leu-gln-phe-ile-arg-glu-thr-glu-glu-ile-val-ser-ala-ser-asn-ser-ser-arg-phe-pro-leu-gly-glu-ser-phe-leu-val-ala-
 2881 AAG-GGA-ATC-CGG-CTG-AGG-AAT-GAG-GAC-CTC-GGC-CTT-CTC-CCG-CTC-TTC-CCG-CCC-CGG-GAG-GCT-TTC-GCG-GAG-TTT-CTG-CGT-GGG-AGT-GAT-
 942 lys-gly-ile-arg-leu-arg-asn-glu-asp-leu-gly-leu-pro-pro-leu-phe-pro-arg-glu-ala-phe-ala-glu-phe-leu-arg-gly-ser-asp-
 2971 TAC-GCC-ATT-CGC-CTG-GCG-GCT-CAG-TCT-ACC-TTA-AGC-TTC-TAT-CAG-AGA-CGC-CGC-TTT-TCC-CCG-GAC-GAC-TGC-GCT-GGA-GCA-TCC-GCC-CTT-
 972 tyr-ala-ile-arg-leu-ala-ala-gln-ser-thr-leu-ser-phe-tyr-gln-arg-arg-arg-phe-ser-pro-asp-asp-ser-ala-gly-ala-ser-ala-leu-
 3061 CTG-CGG-TCG-GGC-CCC-TAC-ATG-CCA-CAG-TGT-GAT-GCG-TTT-GGA-AGT-TGG-GAG-CCT-GTG-CAG-TGC-TAC-GCT-GGG-AGT-GGT-CAC-TGC-TGG-TGT-
 1002 leu-arg-ser-gly-pro-tyr-asn-pro-gln-cys-asp-ala-phe-gly-ser-trp-glu-pro-val-gln-cys-tyr-ala-gly-thr-gly-his-cys-trp-cys-
 3151 GTA-GAT-GAG-AAA-GGA-GGG-TTC-ACT-CCT-GGC-TCA-CTG-ACT-GCC-CGC-TCT-CTG-CAG-ATT-CCA-CAG-TGC-CCG-ACA-ACC-TGC-GAG-AAA-TCT-CGA-
 1032 val-asp-glu-lys-gly-gly-phe-thr-pro-gly-ser-leu-thr-ala-arg-ser-leu-gln-ile-pro-gln-cys-pro-thr-thr-cys-glu-lys-ser-arg-
 3241 ACC-AGT-GGG-CTG-CTT-TCC-AGT-TGG-AAA-CAG-GCT-AGA-TCC-CAA-GAA-AAC-CCA-TCT-CCA-AAA-GAC-CTG-TTC-GTC-CCA-GCC-TGC-CTA-GAA-ACA-
 1062 thr-ser-gly-leu-leu-ser-ser-trp-lys-gln-ala-arg-ser-gln-glu-asn-pro-ser-pro-lys-asp-leu-phe-val-pro-ala-cys-leu-glu-thr-
 3331 GGA-GAA-TAT-GCC-AGG-CTG-CAG-GCA-TCG-GGG-GCT-GGC-ACG-TGG-TGT-GTG-GAC-CCT-GCA-TCA-GGA-GAA-GAG-TTG-CGG-CCT-GGC-TGC-AGC-AGC-
 1092 gly-gly-tyr-ala-arg-leu-gln-ala-ser-gly-ala-gly-thr-trp-cys-val-asp-pro-ala-ser-gly-glu-glu-leu-arg-pro-gly-ser-ser-ser-
 3421 AGT-GCC-CAG-TGC-CCA-AGC-CTC-TGC-AAT-GTG-CTC-AGA-AGT-GGA-GTC-CTC-TCT-AGG-AGA-GTC-AGC-CCA-GGC-TAT-GCT-CCA-GGC-TGC-AGG-GCA-
 1122 ser-ala-gln-cys-pro-ser-leu-cys-asn-val-leu-lys-ser-gly-val-leu-ser-arg-arg-val-ser-pro-gly-tyr-val-pro-ala-cys-arg-ala-
 3511 GAG-GAT-GGG-GGC-TTT-TCC-CCA-GTG-CAA-TGT-GAC-CAG-GCC-CAG-GGC-AGC-TGC-TGG-TGT-GTC-ATG-GAC-AGC-GGA-GAA-GAG-GTG-CCT-GGG-ACG-
 1152 glu-asp-gly-gly-phe-ser-pro-val-gln-cys-asp-gln-ala-gln-gly-ser-cys-trp-cys-val-asn-asp-ser-gly-glu-glu-val-pro-gly-thr-
 3601 CCG-GTG-ACC-GGG-GGC-CAG-CCC-GCC-TGT-GAG-AGC-CCG-CGG-TGT-CCG-CTG-CCA-TTC-AAC-GCG-TCG-GAG-GTG-GTT-GGT-GGA-ACA-ATC-CTG-TGT-
 1182 arg-val-thr-gly-gly-gln-pro-ala-cys-glu-ser-pro-arg-cys-pro-leu-pro-phe-asn-ala-ser-glu-val-val-gly-gly-thr-ile-leu-cys-
 3691 GAG-ACA-ATC-TCG-GGC-CCC-ACA-GGC-TCT-GCC-ATG-CAG-CAG-TGC-CAA-TTG-CTG-TGC-CGC-CAA-GGC-TCC-TGG-AGC-GTG-TTT-CCA-CCA-GGG-CCA-
 1212 glu-thr-ile-ser-gly-pro-thr-gly-ser-ala-asn-gln-gln-cys-gln-leu-leu-cys-arg-gln-gly-ser-trp-ser-val-phe-pro-pro-gly-pro-
 3781 TTG-ATA-TGT-AGC-CTG-GAG-AGC-GGA-CGC-TGG-GAG-TCA-CAG-CTG-CCT-CAG-CCC-CGG-GCC-TGC-CAA-CGG-CCC-CAG-CTG-TGG-CAG-ACC-ATC-CAG-
 1242 leu-ile-cys-ser-leu-glu-ser-gly-arg-trp-glu-ser-gln-leu-pro-gln-pro-arg-ala-cys-gln-arg-pro-gln-leu-trp-gln-thr-gln-ile-gln-
 3871 ACC-CAA-GGG-CAC-TTT-CAG-CTC-CAG-CTC-CCG-CCG-GGC-AGG-ATG-TGC-AGT-GCT-GAC-TAC-GCG-GGT-TTG-CTG-CAG-ACT-TTC-CAG-GTT-TTC-ATA-
 1272 thr-gln-gly-his-phe-gln-leu-gln-leu-pro-pro-gly-lys-asn-cys-ser-ala-asp-tyr-ala-gly-leu-leu-gln-thr-phe-gln-val-phe-ile-
 3961 TTG-GAT-GAG-CTG-ACA-GCC-CGC-GGC-TTC-TGC-CAG-ATC-CAG-GTG-ARG-ACT-TTT-GGC-ACC-CTG-GTT-TCC-ATT-CCT-GTC-TGC-AAC-AAC-TCC-TCT-
 1302 leu-asp-glu-leu-thr-ala-arg-gly-phe-cys-gln-ile-gln-val-lys-thr-phe-gly-thr-leu-val-ser-ile-pro-val-cys-asn-asn-ser-ser-
 4051 GTG-CAG-GTG-GGT-TGT-CTG-ACC-AGG-GAG-CGT-TTA-GGA-GTG-AGT-GTT-ACA-TGG-AAA-TCA-CGG-CIT-GAG-GAC-ATC-CCA-GTG-GCT-TCT-CYT-CCT-
 1332 val-gln-val-gly-cys-leu-thr-arg-glu-arg-leu-gly-val-asn-val-thr-trp-lys-ser-arg-leu-glu-asp-ile-pro-val-ala-ser-leu-pro-
 4141 GAC-TTA-CAT-GAC-ATT-GAG-AGA-GCC-TTG-GTG-GGC-AGG-GAT-CTC-CYT-GGG-CGC-TTC-ACA-GAT-CTG-ATC-CAG-AGT-GGC-TCA-TTC-CAG-CYT-CAT-
 1362 arg-ile-cys-asp-ile-glu-arg-ala-leu-val-gly-lys-asp-leu-leu-gly-arg-phe-thr-asp-leu-ile-gln-ser-gly-ser-phe-gln-leu-his-
 4231 CTG-GAC-TCC-AGA-ACG-TTC-CCA-CGG-GAA-ACC-ATC-CGC-TTC-CTC-CAA-GGG-GAC-CAC-TTT-GGC-ACC-TCT-CCT-AGG-ACA-CGG-TTT-GGG-TGC-TCG-
 1392 leu-asp-ser-lys-thr-phe-pro-ala-glu-thr-ile-arg-phe-leu-gln-gly-asp-his-phe-gly-thr-ser-pro-arg-thr-arg-phe-gly-cys-ser-
 4321 GAA-GGA-TTC-TAC-CAA-GTC-TTG-ACA-AGT-GAG-GCC-AGT-CAG-GAC-GGA-CTG-GGA-TGC-GTT-AGG-TGC-CAT-GAA-GGA-AGC-TAT-TCC-CAA-GAT-GAG-
 1422 glu-gly-phe-tyr-gln-val-leu-thr-ser-glu-ala-ser-gln-asp-gly-leu-gly-cys-val-lys-cys-his-glu-gly-ser-tyr-ser-gln-asp-glu-
 4411 GAA-TGC-ATT-CCT-TGT-CCT-GTT-GGA-TTC-TAC-CAA-GAA-CAG-GCA-GGG-AGC-TTG-GCC-TGT-GTC-CCA-TGT-CCT-GTG-GGC-AGA-ACG-ACC-ATT-TCT-
 1452 glu-cys-ile-pro-cys-pro-val-gly-phe-tyr-gln-glu-gln-ala-gly-ser-leu-ala-cys-val-pro-cys-pro-val-gly-arg-thr-thr-ile-ser-
 4501 GCC-GGA-GCT-TTC-AGC-CAG-ACT-CAC-TGT-GTC-ACT-GAC-TGT-CAG-AGG-ACC-GAA-GCA-GGC-CTG-CAA-TGT-GAC-CAG-AAT-GGC-CAG-TAT-CGA-GCC-
 1482 ala-gly-ala-phe-ser-gln-thr-his-cys-val-thr-asp-cys-gln-arg-asn-glu-ala-gly-leu-gln-cys-asp-gln-asn-gly-gln-tyr-arg-ala-
 4591 ACG-CAG-ARG-GAC-AGG-GGC-AGT-GGG-AGG-GCC-TTC-TGT-ATG-GAC-GGC-GAG-GGG-CGG-AGG-CTG-CCA-TGG-TGG-GAA-ACA-ARG-GCC-CCT-CTT-GAG-
 1512 ser-gln-lys-asp-arg-gly-ser-gly-lys-ala-phe-cys-val-asp-gly-glu-gly-arg-arg-leu-pro-trp-trp-glu-thr-glu-ala-pro-leu-glu-
 4681 GAC-TCA-CAG-TGT-TTG-ATG-ATG-CAG-AGG-TTT-GAG-AGG-GTT-CCA-GAA-TCA-AGG-GTG-ATC-TTC-GAC-GCC-AAT-GCT-CCT-GTG-GCT-GTC-AGA-TCC-
 1542 asp-ser-gln-cys-leu-asn-asn-gln-lys-phe-glu-lys-val-pro-glu-ser-lys-val-ile-phe-asp-ala-asn-ala-pro-val-ala-val-arg-ser-
 4771 AAA-GTT-CCT-GAT-TCT-CAG-TTC-CCC-GTG-ATG-CAG-TGC-TTG-ACA-GAT-TGC-ACA-GAG-GAC-GAG-GCC-TGC-AGC-TTC-TCC-ACC-GTG-TCC-ACG-ACG-
 1572 lys-val-pro-asp-ser-glu-phe-pro-val-asn-gln-cys-leu-thr-asp-cys-thr-glu-asp-glu-ala-cys-ser-phe-phe-thr-val-ser-thr-thr-
 4861 GAG-CCA-GAG-ATT-TCC-TGT-GAT-TTC-TAT-GCT-TGG-ACA-AGT-GAC-AAT-GTT-GCC-TGC-ATG-ACT-TCT-GAC-CAG-AAA-GCA-GAT-GCA-CTG-GGG-ACC-
 1602 glu-pro-glu-ile-ser-cys-asp-phe-tyr-ala-trp-thr-ser-asp-asn-val-ala-cys-asn-thr-ser-asp-gln-lys-arg-asp-ala-leu-gly-asn-
 4951 TCA-AGG-GCC-ACC-AGC-TTT-GGA-AGT-CTT-CGC-TGC-CAG-GTG-AAA-GTG-AGG-AGC-CAT-GGT-CAR-GAT-TCT-CCA-GCT-GTG-TAT-TTG-AAA-ARG-GGC-
 1632 ser-lys-ala-thr-ser-phe-gly-ser-leu-arg-cys-gln-val-lys-val-arg-ser-his-gly-gln-asp-ser-pro-ala-val-tyr-leu-lys-lys-gly-
 5041 CAA-GCA-TCC-ACC-ACA-ACA-CYT-CAG-AAA-CGC-TTT-GAA-CCC-ACT-GGT-TTC-CAA-AAC-ATG-CYT-TCT-GGA-TTG-TAC-AAC-CCC-ATT-GTG-TTC-TCA-
 1662 gln-gly-ser-thr-thr-thr-leu-gln-lys-arg-phe-glu-pro-thr-gly-phe-gln-asn-asn-leu-ser-gly-leu-tyr-asn-pro-ile-val-phe-ser-
 5131 GCC-TCA-GGA-GCC-AAT-CTA-ACC-GAT-GCT-CAC-CTC-TTC-TGT-CYT-TCT-GCA-TGC-GAC-CGT-GAT-CTG-TGT-TGC-GAT-GGC-TTC-CTC-CTC-ACA-CAG-
 1692 ala-ser-gly-ala-asn-leu-thr-asp-ala-his-leu-phe-cys-leu-leu-ala-cys-asp-arg-asp-leu-cys-cys-asp-gly-phe-val-leu-thr-gln-
 5221 GTT-CAA-GGA-GGT-GCC-ATC-ATC-TGT-GGG-TTG-CTG-AGC-TCA-CCC-AGT-GTC-CTG-CYT-TGT-AAT-GTC-AAA-GAC-TGG-ATG-GAT-CCC-TCT-GAA-GCC-
 1722 val-gln-gly-gly-ala-ile-ile-cys-gly-leu-leu-ser-ser-pro-ser-val-leu-leu-cys-asn-val-lys-asp-trp-asn-asp-pro-ser-glu-glu-
 5311 TGG-GCT-AAT-GCT-ACA-TGT-CCT-GGT-GTG-ACA-TAT-GAC-CAG-GAG-AGC-CAC-CAG-GTG-ATA-TTG-CGT-CTT-GGA-GAC-CAG-GAG-TTC-ATC-ARG-AGT-
 1752 trp-ala-asn-ala-thr-cys-pro-gly-val-thr-tyr-asp-gln-glu-ser-his-gln-val-ile-leu-arg-leu-gly-asp-gln-glu-phe-ile-lys-ser-
 5401 CTO-ACA-CCC-TTA-GAA-AAA-ACT-CAA-GAC-ACC-TTT-ACC-AAT-TTT-CAG-CAG-GTT-TAT-CTC-TGG-AAA-GAT-TCT-GAC-ATG-GGG-TCT-CGG-CCT-GAG-
 1782 thr-thr-pro-leu-glu-gly-thr-gln-asp-thr-phe-thr-asp-phe-gln-gln-val-tyr-leu-trp-lys-asp-ser-asp-asn-gly-ser-arg-pro-glu-
 5491 TCT-ATG-GGA-TGT-AGA-AAA-ACC-ACA-GTG-CCA-AGG-CCA-GCA-TCT-CCA-ACA-GAA-GCA-GGT-TTG-ACA-ACA-GAA-CTT-TTC-CCT-CCT-GTG-GAC-CTC-
 1812 ser-asn-gly-cys-arg-lys-asn-thr-val-pro-arg-pro-ala-ser-pro-thr-glu-ala-gly-leu-thr-thr-glu-leu-phe-ser-pro-val-asp-leu-
 5581 AAC-CAG-GTC-ATT-GTC-AAT-GGA-AAT-CAA-TCA-CTA-TCC-AGC-CAG-ARG-CAC-TGG-CIT-TTC-ARG-CAC-CTG-TTT-TCA-GCC-CAG-GCA-ACC-CTA-
 1842 asn-gln-val-ile-val-asn-gly-asn-gln-ser-leu-ser-ser-gln-lys-his-trp-leu-phe-lys-his-leu-phe-ser-ala-gln-gln-ala-asn-leu-

5671 TGG-TGC-CTT-TCT-CGT-TGT-GTG-CAG-GAG-CAC-TCT-TTC-TGT-CAG-CTC-GCA-GAG-ATA-ACA-GAG-AGT-GCA-TCC-TTG-TAC-TTC-ACC-TGC-ACC-CTC-
 1872 trp-cys-leu-ser-arg-cys-val-gln-glu-his-ser-phe-cys-gln-leu-ala-glu-ile-thr-glu-ser-ala-ser-leu-tyr-phe-thr-cys-thr-leu-
 5761 TAC-CCA-GAG-GCA-CAG-GTG-TGT-GAT-GAC-ATC-ATG-GAG-TCC-AAT-ACC-CAG-GGC-TGC-AGA-CTG-ATC-CTG-CCT-CAG-ATG-CCA-AGG-GCC-CTG-TTC-
 1902 tyr-pro-glu-ala-gln-val-cys-asp-asp-ile-met-glu-ser-asn-thr-gln-gly-cys-arg-leu-ile-leu-pro-gln-met-pro-lys-ala-leu-phe-
 5851 CGG-ARG-AAA-GTT-ATA-CTG-GAA-GAT-AAA-GTG-ARG-AAC-TTT-TAC-ACT-CGC-CTG-CCG-TTC-CAA-AAA-CTG-ATG-GGG-ATA-TCC-ATT-AGA-AAT-AAA-
 1932 arg-lys-lys-val-ile-leu-glu-asp-lys-val-lys-asn-phe-tyr-thr-arg-leu-pro-phe-gln-lys-leu-met-gly-ile-ser-ile-arg-asn-lys-
 5941 GTG-CCC-ATG-TCT-GAA-AAA-TCT-ATT-TCT-AAT-GGG-TTC-TTT-GAA-TGT-GAA-CGA-CGG-TGC-GAT-GCG-GAC-CCA-TGC-TGC-ACT-GGC-TTT-GGA-TTT-
 1962 val-pro-met-ser-glu-lys-ser-ile-ser-asn-gly-phe-phe-glu-cys-glu-arg-arg-cys-asp-ala-asp-pro-cys-cys-thr-gly-phe-gly-phe-
 6031 CTA-AAT-GTT-TCC-CAG-TTA-AAA-GGA-GGA-GAG-GTG-ACA-TGT-CTC-ACT-CTG-AAC-AGC-TTG-GGA-ATT-CAG-ATG-TGC-AGT-GAG-GAG-AAT-GGA-GGA-
 1992 leu-asn-val-ser-gln-leu-lys-gly-gly-glu-val-thr-cys-leu-thr-leu-asn-ser-leu-gly-ile-gln-met-cys-ser-glu-glu-asn-gly-gly-
 6121 GCC-TGG-CGC-ATT-TTG-GAC-TGT-GGC-TCT-CCT-GAC-ATT-GAA-GTC-CAC-ACC-TAT-CCC-TTC-GGA-TGG-TAC-CAG-ARG-CCC-ATT-GCT-CAA-AAT-AAT-
 2022 ala-trp-arg-ile-leu-asp-cys-gly-ser-pro-asp-ile-glu-val-his-thr-tyr-pro-phe-gly-trp-tyr-gln-lys-pro-ile-ala-gln-asn-asn-
 6211 GCT-CCC-AGT-TTT-TGC-CCT-TTG-GTT-GTT-CTG-CCT-TCC-CTC-ACA-GAG-AAA-GTG-TCT-CTG-GAA-TGG-TGG-CAG-TCC-CTG-GCC-CTC-TCT-TCA-GTG-
 2052 ala-pro-ser-phe-cys-pro-leu-val-val-leu-pro-ser-leu-thr-glu-lys-val-ser-leu-glu-ser-trp-gln-ser-leu-ala-leu-ser-ser-val-
 6301 GTT-GTT-GAT-CCA-TCC-ATT-AGG-CAC-TTT-GAT-GTT-GCC-CAT-GTC-AGC-ACT-GCT-GCC-ACC-AGC-AAT-TTC-TCT-GCT-GTC-CGA-GAC-CTC-TGT-TTG-
 2082 val-val-asp-pro-ser-ile-arg-his-phe-asp-val-ala-his-val-ser-thr-ala-ala-thr-ser-asn-phe-ser-ala-val-arg-asp-leu-cys-leu-
 6391 TCG-GAA-TGT-TCC-CAA-CAT-GAG-GCC-TGT-CTC-ATC-ACC-ACT-CTG-CAA-ACC-CAA-CTC-GGG-GCT-GTG-AGA-TGT-ATG-TTC-TAT-GCT-GAT-ACT-CAA-
 2112 ser-glu-cys-ser-gln-his-glu-ala-cys-leu-ile-thr-thr-leu-gln-thr-gln-leu-gly-ala-val-arg-cys-met-phe-tyr-ala-asp-thr-gln-
 6481 AGC-TGC-ACA-CAT-AGT-CTG-CAG-GGT-CGG-AAC-TGC-CGA-CTT-CTG-CIT-CGT-GAA-GAG-GCC-ACC-CAC-ATC-TAC-CGG-ARG-CCA-GGA-ATC-TCT-CTG-
 2142 ser-cys-thr-his-ser-gln-gly-arg-asn-cys-arg-leu-leu-arg-glu-glu-ala-thr-his-ile-tyr-arg-lys-pro-ile-gln-val-gly-thr-ser-
 6571 CTC-AGC-TAT-GAG-CGA-TCT-GTA-CCT-TCT-GTG-CCC-ATT-TCC-ACC-GGC-CGG-CTG-CTG-GGC-AGG-TCC-CAG-GCC-ATC-CAG-GTG-GGT-ACC-TCA-
 2172 leu-ser-tyr-gln-ala-ser-val-pro-ser-val-pro-ile-ser-thr-his-gly-arg-leu-leu-gly-arg-ser-gln-ala-ile-gln-val-gly-thr-ser-
 6661 TGG-ARG-CAA-GTG-GAC-CAG-TTC-CIT-GGA-GTT-CCA-TAT-GCT-GCC-CCG-CCC-CTG-GCA-GAG-AGG-CAC-TTC-CAG-GCA-CCA-GAG-CCC-TTG-AAC-TGG-
 2202 trp-lys-gln-val-asp-gln-phe-leu-gly-val-pro-tyr-ala-ala-pro-pro-leu-ala-glu-arg-his-phe-gln-ala-pro-glu-pro-leu-asn-trp-
 6751 ACA-GGC-TCC-TGG-GAT-GCC-AGC-ARG-CCA-AGG-GCC-AGC-TGC-TGG-CAG-CCA-GGC-ACC-AGA-ACA-TCC-ACG-TCT-CCT-GGA-GTC-AGT-GAA-GAT-TGT-
 2232 thr-gly-ser-trp-asp-ala-ser-lys-pro-arg-ala-ser-cys-trp-gln-pro-gly-thr-arg-thr-ser-thr-ser-pro-gly-val-ser-glu-asp-cys-
 6841 TTG-TAT-CTC-AAT-GTG-TTC-ATC-CCT-CAG-AAT-GTG-GCC-CCT-AAC-GCG-TCT-GTG-CTG-GTG-TTC-TTC-CAC-AAC-ACC-ATG-GAC-AGG-GAG-GAG-AGT-
 2262 leu-tyr-leu-asn-val-phe-ile-pro-gln-asn-val-ala-pro-asn-ala-ser-val-leu-val-phe-phe-his-asn-thr-met-asp-arg-glu-glu-ser-
 6931 GAA-GGA-TGG-CCG-GCT-ATC-GAC-GGC-TCC-TTC-TTG-GCT-GCT-GTT-GGC-AAC-CTC-ATC-GTG-GTC-ACT-GCC-AGC-TAC-CGA-GTG-GGT-GTC-TTC-GGC-
 2292 glu-gly-trp-pro-ala-ile-asp-gly-ser-phe-leu-ala-ala-val-gly-asn-leu-ile-val-val-thr-ala-ser-tyr-arg-val-gly-val-phe-gly-
 7021 TTC-CTG-AGT-TCT-GGA-TCC-GGA-GAG-GTG-AGT-GGC-AAC-TGG-GGG-CTG-CTG-GAC-CAG-GTG-GCG-GCT-CTG-ACC-TGG-GTG-CAG-ACC-CAC-ATC-CGA-
 2322 phe-leu-ser-ser-gly-ser-gly-glu-val-ser-gly-asn-trp-gly-leu-leu-asp-gln-val-ala-ala-leu-thr-trp-val-gln-thr-his-ile-arg-
 7111 GGA-TTT-GGC-GGG-GAC-CCT-CGG-CGC-GTG-TCC-CTG-GCA-GCA-GAC-CGT-GGC-GGG-GCT-GAT-GTG-GCC-AGC-ATC-CAC-CTT-CTC-ACG-GCC-AGG-GCC-
 2352 gly-phe-gly-gly-asp-pro-arg-arg-val-ser-leu-ala-ala-asp-arg-gly-gly-ala-asp-val-ala-ser-ile-his-leu-leu-thr-ala-arg-ala-
 7201 ACC-AAC-TCC-CAA-CTT-TTC-CGG-AGA-GCT-GTG-CTG-ATG-GGA-GGC-TCC-GCA-CTC-TCC-CCG-GCC-GCC-GTC-ATC-AGC-CAT-GAG-AGG-GCT-CAG-CAG-
 2382 thr-asn-ser-leu-phe-arg-arg-ala-val-leu-met-gly-gly-ser-ala-leu-ser-pro-ala-ala-val-ile-ser-his-glu-arg-ala-gln-gln-
 7291 CAG-GCA-ATT-GCT-TTG-GCA-ARG-GAG-GTC-AGT-TGC-CCC-ATG-TCA-TCC-AGC-CAA-GAA-GTG-GTG-TCC-TGC-CTC-CGC-CAG-ARG-CCT-GCC-AAT-GTC-
 2412 gln-ala-ile-ala-leu-ala-lys-glu-val-ser-cys-pro-met-ser-ser-ser-gln-glu-val-val-ser-cys-leu-arg-gln-lys-pro-ala-asn-val-
 7381 CTC-AAT-GAT-GCC-CAG-ACC-ARG-CTC-CTG-GCC-GTG-AGT-GGC-CCT-TTC-CAC-TAC-TGG-GGT-CCT-GTG-ATC-GAT-GGC-CAC-TTC-CTC-CGT-GAG-CCT-
 2442 leu-asn-asp-ala-gln-thr-lys-leu-leu-ala-val-ser-gly-pro-phe-his-tyr-trp-gly-pro-val-ile-asp-gly-his-phe-leu-arg-glu-pro-
 7471 CCA-GCC-AGA-GCA-CTG-ARG-AGG-TCT-TTA-TGG-GTA-GAG-GTC-GAT-CTG-CTC-ATT-GGG-AGT-TCT-CAG-GAC-GAC-GGG-CTC-ATC-AAC-AGA-GCA-ARG-
 2472 pro-ala-arg-ala-leu-lys-arg-ser-leu-trp-val-glu-val-asp-leu-leu-ile-gly-ser-ser-gln-asp-asp-gly-leu-ile-asn-arg-ala-lys-
 7561 GCT-GTG-ARG-CAA-TTT-CAG-GAA-AGT-CGA-GGC-CGG-ACC-AGT-AGC-AAA-ACA-GCC-TTT-TAC-CAG-GCA-CTG-CAG-AAT-TCT-CTG-GGT-GGC-GAG-GAC-
 2502 ala-val-lys-gln-phe-glu-glu-ser-arg-gly-arg-thr-ser-ser-lys-thr-ala-phe-tyr-gln-ala-leu-gln-asn-ser-leu-gly-gly-glu-asp-
 7651 TCA-GAT-GCC-CGC-GTC-CAG-GCT-GCT-GCT-ACA-TGG-TAT-TAC-TCT-CTG-GAG-CAC-TCC-ACG-GAT-GAC-TAT-GCC-TCC-TTC-TCC-CGG-GCT-CTG-GAG-
 2532 ser-asp-ala-arg-val-glu-ala-ala-ala-thr-trp-tyr-tyr-ser-leu-glu-his-ser-thr-asp-asp-tyr-ala-ser-phe-ser-arg-ala-leu-glu-
 7741 AAT-GCC-ACC-CGG-GAC-TAC-TTT-ATC-ATC-TGC-CCT-ATA-ATC-GAC-ATG-GCC-AGT-GCC-TGG-GCA-ARG-AGG-GCC-CGA-GGA-AAC-GTC-TTC-ATG-TAC-
 2562 asn-ala-thr-arg-asp-tyr-phe-ile-ile-cys-pro-ile-ile-asp-met-ala-ser-ala-trp-ala-lys-arg-ala-arg-gly-asn-val-phe-met-tyr-
 7831 CAT-GCT-CCT-GAA-AAC-TAC-GGC-CAT-GGC-AGC-CTG-CAG-CTG-CTG-GCG-GAT-GTT-CAG-TTT-GCC-TTG-GGG-CYT-CCC-TTC-TAC-CCA-GCC-TAC-CAG-
 2592 his-ala-pro-glu-asn-tyr-gly-his-gly-ser-leu-glu-leu-leu-ala-asp-val-gln-phe-ala-leu-gly-leu-pro-phe-tyr-pro-ala-tyr-glu-
 7921 GGG-CAG-TTT-TCT-CTG-GAG-GAG-ARG-AGC-CTG-CTG-AAA-ATC-ATG-CAG-TAC-TTT-TCC-CAC-TTC-ATC-AGA-TCA-GGA-AAT-CCC-AAC-TAC-CCT-
 2622 gly-ser-leu-leu-glu-gly-ser-leu-ser-leu-lys-ile-met-gln-tyr-phe-ser-his-phe-ile-arg-ser-gly-asn-pro-asn-tyr-pro-
 8011 TAT-GAG-TTC-TCA-CGG-AAA-GTA-CCC-ACA-TTT-GCA-ACC-CCC-TGG-CCT-GAC-TTT-GTA-CCC-CGT-GCT-GGT-GGA-GAG-AAC-TAC-ARG-GAG-TTC-AGT-
 2652 tyr-glu-phe-ser-arg-lys-val-pro-thr-phe-ala-thr-pro-trp-pro-asp-phe-val-pro-arg-ala-gly-gly-glu-asn-tyr-lys-glu-phe-ser-
 8101 GAG-CTG-CTC-CCC-AAT-CCA-CAG-GGC-CTG-ARG-AAA-GCC-TGC-TCC-TTC-TGG-TCC-ARG-TAC-ATC-TCG-TCT-CTG-ARG-ACA-TCT-GCA-GAT-GGA-
 2682 glu-leu-leu-pro-asn-arg-gln-gly-leu-lys-lys-ala-asp-cys-ser-phe-trp-ser-lys-tyr-ile-ser-ser-leu-lys-thr-ser-ala-asp-gly-
 8191 GCC-ARG-GGC-GGG-CAG-TCA-GCA-GAG-AGT-GAA-GAG-GAG-TTG-ACG-GCT-GGA-TCT-GGG-CTA-AGA-GAA-GAT-CTC-CTA-AGC-CTC-CAG-GAA-CCA-
 2712 ala-lys-gly-gly-gln-ser-ala-glu-ser-glu-glu-glu-glu-thr-ala-gly-ser-gly-leu-arg-glu-asp-leu-leu-ser-leu-gln-glu-pro-
 8281 GGC-TCT-ARG-ACC-TAC-ACC-ARG-TGA-CCAGCCCTTGACCTCCCAAAACCTACCCAGGCTGCCACTATGGTCTATCTTTTCTCTAAATAGTACTTACCTTCATAA-
 2742 gly-ser-lys-thr-tyr-ser-lys
 TATCTACATCCGGT

Fig. 2

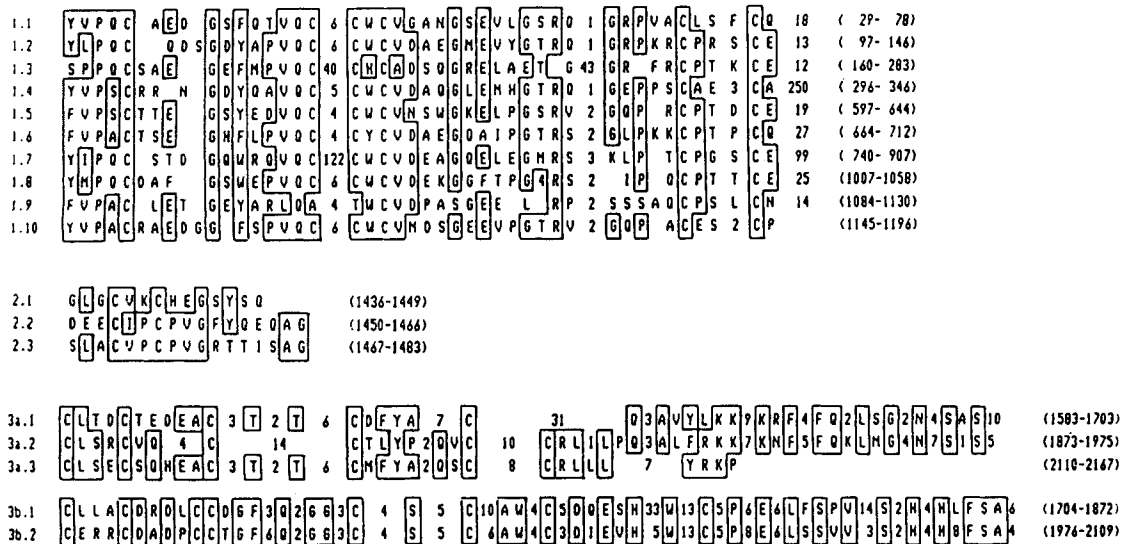


Fig. 3. Internal homologies in human thyroglobulin primary structure. The three types of homologous domains are aligned. Insertions or deletions have been introduced to optimize homology. Conserved amino acids are boxed according to the following scheme: for type 1 repeats, the amino acids should be present at least five times; for type 2 and type 3 repeats, at least twice. Residues showing no conservation between the domains are represented by numbers. Localization of the domains is shown on the right (the numbering is as in Fig. 2)

Table 1. The amino acid composition of the human thyroglobulin monomer

Amino acid	Number of residues	Percentage
A Alanine	195	7.10
R Arginine	151	5.50
D Aspartic acid	124	4.51
N Asparagine	83	3.02
C Cysteine	122	4.44
E Glutamic acid	169	6.15
Q Glutamine	191	6.95
G Glycine	209	7.60
H Histidine	41	1.49
I Isoleucine	81	2.95
L Leucine	258	9.39
K Lysine	91	3.31
M Methionine	37	1.35
F Phenylalanine	140	5.09
P Proline	171	6.22
S Serine	267	9.72
T Threonine	146	5.31
W Tryptophan	43	1.56
Y Tyrosine	67	2.44
V Valine	162	5.90

of sites actually glycosylated in the mature protein. One of them was actually displayed in human thyroglobulin, at position 2562 [25].

Analysis of the deduced amino-acid sequence of the protomer reveals the existence of three types of repeated domains. Type 1 is exclusively present in the N-terminal portion of the molecule, and is repeated ten times between positions 29 and 1196. It is composed of approximately 50 amino acids, in which the positions of Cys, Pro and Gly residues are highly conserved. The proportion of Cys and Tyr in the type 1 domains is high, as compared to the entire protomer. Some insertions of variable length are found in fixed positions in the area, but seem to have no relation with the repeated domains. The second repeated domains is shorter (14–17 residues) and is present in triplicate between positions 1436 and 1438. The third domain (type 3) is repeated five times, between positions 1603 and 2186. As in type 1, maximal

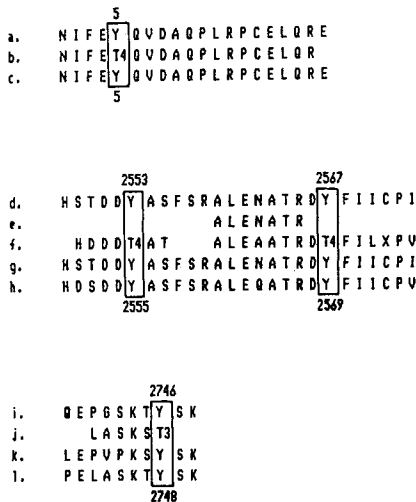


Fig. 4. Hormonogenic domains in human thyroglobulin. The hormone-containing domains identified to date in various species are represented by the one-letter code and are aligned for optimal homology. The four hormonogenic tyrosines are boxed and their positions in the protein are indicated. a, d, i, peptide sequences derived from the present work; b, four identical hormone-containing peptides determined from human, bovine, ovine and porcine thyroglobulins [33, 34, 12, 35]; c, h, l, protein sequences derived from the bovine cDNA sequence [28]; the C-terminal region of the ovine thyroglobulin is identical to that in sequence 1 (Durand, J. and Malhiéry, Y., unpublished); g, k, protein sequence derived from a partial rat cDNA sequence [36]; f, sequence of three peptides obtained from porcine thyroglobulin [10]; i, sequence of a hormone-containing peptide isolated from porcine thyroglobulin [11]; e, tryptic peptide, isolated from human thyroglobulin, containing a glycosidic chain attached to the asparagine residue [25]

homology is observed near the Cys residues, whose positions are highly conserved. A more refined analysis of the type 3 domains allows one to distinguish two subtypes (3a and 3b, see Fig. 3), suggesting that type 3 region originates from tandem duplication of an ancestral gene segment.

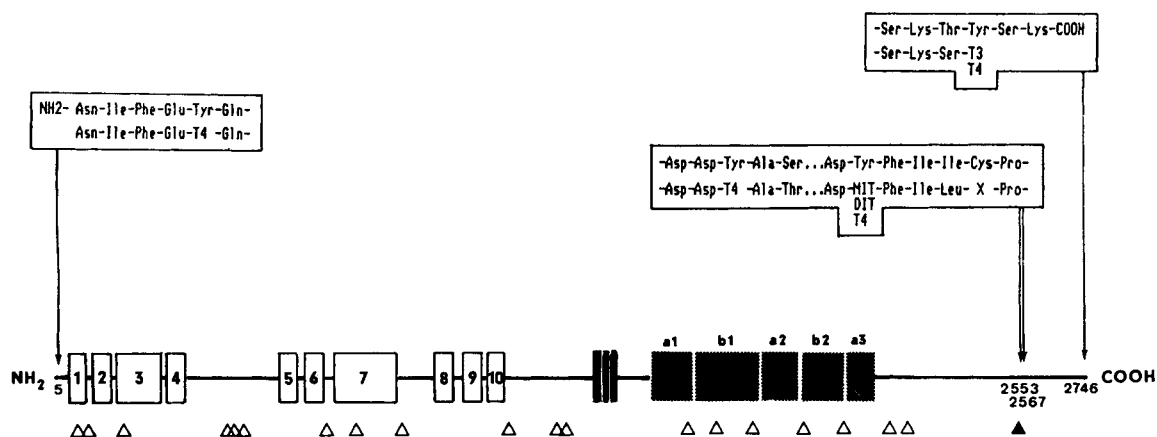


Fig. 5. Schematic representation of the human thyroglobulin monomer. The ten domains of the type 1 homology are represented by open boxes. The three domains of the type 2 homology are represented in black, in the middle of the molecule. The type 3 homology (hatched boxes) is constituted of subtypes 3a (a1, a2, a3) and 3b (b1, b2). The triangles correspond to potential glycosylation sites; the closed triangle corresponds to the actual glycosylation site (see Fig. 4). The four hormonogenic sites are indicated by arrows; in the boxes, the upper line shows the sequence deduced from the nucleotide sequence, the bottom line corresponds to the peptide sequence characterized on the mature human or porcine proteins (see Fig. 4). MIT = monoiodotyrosine; DIT = diiodotyrosine; T3 = 3,5,3'-triiodothyronine; T4 = thyroxine

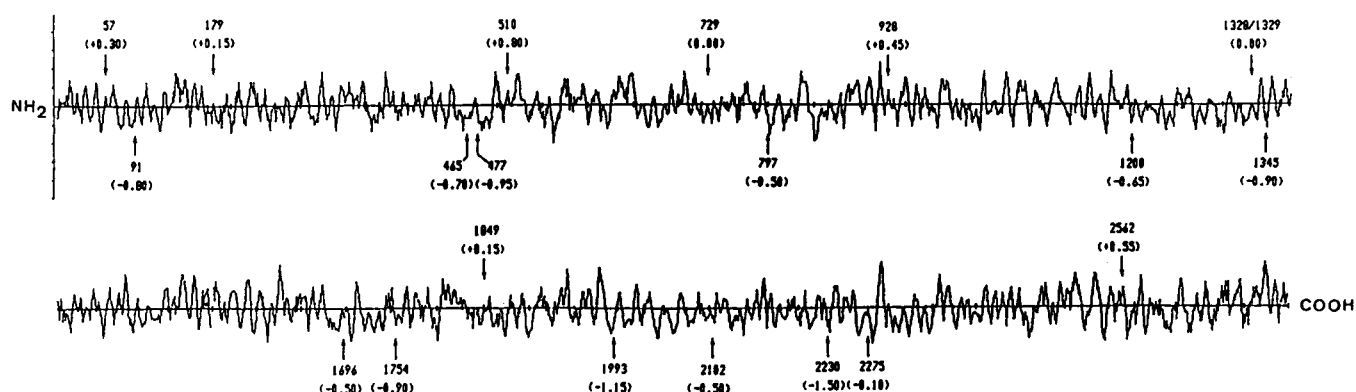


Fig. 6. Hydrophilicity profile of the human thyroglobulin monomer, starting from the N-terminal of the molecule. The computer program used and the hydrophilicity coefficients of the amino acids are from Hopp and Woods [26], using a five-amino-acid window. Figures indicate the positions of the potential N-glycosylation sites (Asn residues in the tripeptide Asn-Xaa-Ser/Thr) and the corresponding hydrophilicity indexes, taken as a measure of the hydrophilicity of the environment of that amino acid [26]

DISCUSSION

The analysis of the amino-acid sequence of human thyroglobulin, as deduced from the complete sequence of its mRNA, reveals that the 2767 amino acid protomer consists of two regions with distinct structural characteristics. Within the 2200 amino-terminal residues of the protein, three families of homologous domains could be detected, for a total of 18 repeats (Fig. 5). By contrast, no internal homologies could be observed within the 580 C-terminal residues. A low cysteine content and a high tyrosine proportion (up to 9% of total residues between the positions 2539 and 2671) further distinguishes the C-terminal from the rest of the protomer. Those structural differences might be related to functional differences, since Marriq et al. [10] and Rawitch et al. [12] demonstrated that four of the tyrosine residues involved as 'acceptor' sites in thyroxine and triiodothyronine formation were situated in positions 5, 2553, 2567 and 2746 of the mature protein (Fig. 4). If the central portion of the molecule does not contain hormonogenic sites, it could play a structural role, either in the positioning of the hormonogenic tyrosines, or in

the endocytosis of thyroglobulin. Unfortunately, analysis of variations in hydrophobicity along the protomer, performed according to Hopp and Woods [26], showed no important differences between the three domains. It is noteworthy however that the only N-glycosylation site whose position has been determined (2562) is part of a hydrophilic domain (Fig. 6). Since approximately half of the 20 potential glycosylation sites are actually glycosylated in the mature protein, and since 8 sites out of 20 potential residues show a positive hydrophilicity index, the search for glycosylated residues should begin with these sites [27].

An organisation similar to that of human thyroglobulin was recently reported for bovine thyroglobulin [28], including the same number of internal repeats in the central portion of the molecule. The bovine mRNA is slightly shorter, but the corresponding protomer is two amino acids longer. Actually, the bovine protomer shows six single amino-acid alterations, as compared to the human protomer, two deletions in positions 440 and 1799, and four insertions in positions 792, 985, 1422 and 1450. Homology between the two mRNAs is 81.25% (81.75% on the coding region) and 77.22% between

the protomers, but increases in the hormonogenic regions, up to 100% for the 20 N-terminal residues (see Fig. 4). The strong homology between human and bovine thyroglobulin mRNAs, together with data on the structure of the human thyroglobulin gene suggest that the three domains observed in the mRNA structure might be evolutionally different. In the central region, homologies of types 1 and 3 probably originate from multiple duplications.

Computer searching in a data bank for homologies between the type 1 repeat and fragments of other proteins revealed that a tripeptide Cys-Trp/Tyr-Cys, whose position is highly conserved in all 10 cases of the repeat, was found in all known scorpion neurotoxins. Neurotoxins are short polypeptides of about 60 amino-acid residues [29], in which the tripeptide helps to maintain a strict tridimensional structure, but its role in thyroglobulin is unknown. Furthermore, the distribution of the other cysteine residues in the type 1 repeat and in neurotoxins is similar. More intriguing is the homology found between the C-terminal (non-repetitive) end of thyroglobulin and *Torpedo californica* acetylcholinesterase, as described by Schumacher et al. [30]. The amazing 64% homology between segments 2314–2360 of human thyroglobulin and 147–197 of acetylcholinesterase is suggestive of a common function that was conserved during evolution. That function is unknown as yet, although several hypotheses have been put forward [31].

Knowledge of the structure of human thyroglobulin mRNA, and of the organisation of the corresponding gene [32], should facilitate the development of studies on the structural bases of defects in thyroglobulin production. Along this line, defect in the structure of bovine thyroglobulin mRNA has already been linked to the existence of a hereditary goitre [13].

We are grateful to J. C. Dagorn for helpful discussion and typing of the manuscript, and to G. de Lanversin for helpful computer analysis. This work was supported by grants from the *Centre National de la Recherche Scientifique* (U.A. 178 and *Action Thématique Programmée*) and *Institut National de la Santé et de la Recherche Médicale* (U.38).

REFERENCES

1. Van Herle, A. J., Vassart, G. & Dumont, J. E. (1979) *New Engl. J. Med.* **301**, 239–249 & 307–314.
2. Baumann, R. & Ross, E. (1895) *Z. Physiol. Chem.* **21**, 281.
3. Lissitzky, S., Mauchamp, J., Reynaud, J. & Rolland, M. (1975) *FEBS Lett.* **60**, 359.
4. Edelhofer, H. & Robbins, J. (1978) *The thyroid* (Werner, S. C. & Ingbar, S. H. eds) pp. 62–76, Harper and Row, New York.
5. Chebath, J., Chabaud, O., Bergé-Lefranc, J. L., Cartouzou, G. & Lissitzky, S. (1977) *Biochem. Biophys. Res. Commun.* **79**, 267–273.
6. Vassart, G., Verstreken, L. & Dinsart, C. (1977) *FEBS Lett.* **79**, 15–18.
7. Bergé-Lefranc, J. L., Cartouzou, G., Malthiery, Y., Perrin, F., Jarry, B. & Lissitzky, S. (1981) *Eur. J. Biochem.* **120**, 1–7.
8. Smeds, S. (1970) *Biochem. Biophys. Res. Commun.* **38**, 1168.
9. Rolland, M., Montfort, M. F., Valenta, L. & Lissitzky, S. (1972) *Clin. Chim. Acta* **39**, 95.
10. Marriq, C., Rolland, M. & Lissitzky, S. (1982) *EMBO J.* **1**, 397–401.
11. Marriq, C., Rolland, M. & Lissitzky, S. (1983) *Biochem. Biophys. Res. Commun.* **112**, 206–213.
12. Rawitch, A. B., Chernoff, St. B., Litwa, M. R., Rouse, J. B. & Hamilton, J. W. (1983) *J. Biol. Chem.* **258**, 2079–2082.
13. Ricketts, M. H., Pohl, V., de Martynoff, G., Boyd, C. D., Bester, A. J., Van Jaarsved, P. P. & Vassart, G. (1985) *EMBO J.* **4**, 731–737.
14. Brocas, H., Christophe, D., Pohl, V. & Vassart, G. (1982) *FEBS Lett.* **137**, 189–192.
15. Maniatis, T., Fritsch, E. F. & Sambrook, J. (1982) *Molecular cloning, a laboratory manual*, Cold Spring Harbor Laboratory, New York.
16. Maxam, A. & Gilbert, W. (1980) *Methods Enzymol.* **65**, 499–560.
17. Alvino, C. G., Tassi, V., Paterson, B. M. & Di Lauro, R. (1982) *FEBS Lett.* **134**, 307–313.
18. Glisin, V., Crkvenjakov, R. & Byus, C. (1974) *Biochemistry* **13**, 2633.
19. Malthiery, Y. & Lissitzky, S. (1985) *Eur. J. Biochem.* **147**, 53–58.
20. Malthiery, B., Bellon, B., Giorgi, D. & Jacq, B. (1984) *Nucleic Acids Res.* **12**, 569–579.
21. Christophe, D., Cabrer, B., Baccola, A., Targovnik, H., Pohl, V. & Vassart, G. (1985) *Nucleic Acids Res.* **13**, 5127–5144.
22. Chebath, J., Chabaud, O., Becarevic, A., Cartouzou, G. & Lissitzky, S. (1977) *Eur. J. Biochem.* **77**, 243.
23. Salvatore, G., Salvatore, M., Cahnmann, H. & Robbins, J. (1964) *J. Biol. Chem.* **239**, 3267.
24. Rolland, M., Bismuth, J., Fondarai, J. & Lissitzky, S. (1966) *Acta Endocrinol.* **53**, 286.
25. Rawitch, A. B., Lias, T. H. & Pierce, J. G. (1968) *Biochim. Biophys. Acta* **160**, 360–367.
26. Hopp, T. P. & Woods, K. R. (1981) *Proc. Natl Acad. Sci. USA* **78**, 3824–3828.
27. Kornfeld, R. & Kornfeld, S. (1985) *Annu. Rev. Biochem.* **54**, 631–664.
28. Mercken, L., Simmons, M. J., Swillens, S., Massaer, M. & Vassart, G. (1985) *Nature (Lond.)* **316**, 647–651.
29. Rochat, H., Bernard, P. & Couraud, F. (1979) *Adv. Cytopharmacol.* **3**, 325–333.
30. Schumacher, M., Camp, S., Maulet, Y., Newton, M., Mac Phee Quigley, K., Taylor, S. S., Friedmann, T. & Taylor, P. (1986) *Nature (Lond.)* **319**, 407–409.
31. Swillens, S., Ludgate, M., Mercken, L., Dumont, J. E. & Vassart, G. (1986) *Biochem. Biophys. Res. Commun.* **137**, 142–148.
32. Baas, F., Van Ommen, G. J. B., Bikker, H., Arnberg, A. C. & de Vijlder, J. J. M. (1986) *Nucleic Acids Res.* **14**, 5171–5186.
33. Lejeune, P. J., Marriq, C., Rolland, M. & Lissitzky, S. (1983) *Biochem. Biophys. Res. Commun.* **114**, 73–80.
34. Marriq, C., Lejeune, P. J., Malthiery, Y., Rolland, M. & Lissitzky, S. (1984) *FEBS Lett.* **175**, 140–146.
35. Rawitch, A. B., Litwa, M. R., Gregg, J., Turner, C. D., Rouse, J. B. & Hamilton, J. W. (1984) *Biochem. Biophys. Res. Commun.* **118**, 423–429.
36. Di Lauro, R., Obici, S., Condcliffe, D., Ursini, M., Musti, A., Moscatelli, C. & Avvedimento, V. E. (1985) *Eur. J. Biochem.* **148**, 7–11.