New insights in the human thyroglobulin structure
van de Graaf, S.A.R.
Alternative polyadenylation cleavage sites in human thyroglobulin mRNA.

E Pauws, SAR van de Graaf, JJM de Vlijder, C Ris-Stalpers.

ABSTRACT
After constructing and analyzing a SAGE (Serial Analysis of Gene Expression) expression library, 6099 different mRNA molecules from a human normal thyroid gland were identified. A SAGE tag corresponding to the mRNA of thyroglobulin (Tg), coding for the most abundant thyroid-specific protein involved in thyroid hormone synthesis, was present. This 10bp tag was located in the 3' end of the mRNA spanning the polyadenylation cleavage site. Apart from the 'wild-type' tag, 2 other tags were present in the library at a lower expression level closely resembling the Tg tag sequence, but could not be identified as such through GenBank. To study if these tags represented Tg mRNA molecules with alternative cleavage sites a total of 62 Tg-3' RACE clones were sequenced. The results show that the 3 putative Tg SAGE tags must be attributed to Tg transcripts and reflect the use of alternative polyadenylation cleavage sites in vivo. In addition, an extra polyadenylation variant of Tg was found with the cleavage site 16bp downstream of the wild-type polyA-site. All cleavage sites match the consensus for polyadenylation cleavage sites in eukaryotic RNA processing. This evidence of alternative polyadenylation within a single mRNA implies a biological inaccuracy of the polyadenylation machinery in vivo.
INTRODUCTION

Polyadenylation of eukaryotic mRNAs is characterized by the cleavage of the precursor-RNA and the addition of a poly(A)tail. Polyadenylation of a primary RNA transcript is suggested to have a function in RNA metabolism such as export of the mature mRNA, stability and recognition by ribosomes. The signals that determine the exact site of polyadenylation have been studied extensively in in vitro experiments. The upstream element AAUAAA together with a downstream U/GU-rich element are present in nearly all eukaryotic (pre)mRNAs and are proposed to regulate the complex machinery of proteins necessary to complete polyadenylation; for review see (1-3). The exact site of cleavage in the pre-mRNA is determined by the position of the regulatory elements together with the specific nucleotide composition in the cleavage region (fig. 1). This approximately 6bp region is typically 11-24bp downstream from a AAUAAA element and 10-30bp upstream from a U/GU-rich element. The preferred nucleotide at the cleavage site is ordered A>U>C>G (4). In general, in a given mRNA sequence one specific cleavage site is used by the polyadenylation machinery although mutational analysis shows that in in vitro cleavage experiments neighboring nucleotides can be used (5,6). The enzyme responsible for the actual cleavage of the mRNA molecule and its mechanism is until now unknown.

Thyroglobulin (Tg) is the most abundant protein present in the thyroid gland functioning as a matrix protein for the synthesis and storage of thyroid hormone (7). In studying the expression profile of the human thyroid we constructed a SAGE (Serial Analysis of Gene Expression) library containing 10994 sequence tags each representing a mRNA molecule. In total 6099 different mRNAs and their relative abundance was determined. Tg was found to be the highest expressed mRNA in the thyroid with a relative abundance of 2.6% in the total mRNA pool (8). The principle of
Chapter 3

the SAGE technique is that each 10bp sequence tag corresponds to a unique mRNA considering its position 3' to the last NsilI restriction site in the cDNA (9). In the case of Tg the so-called SAGE tag is located at the very end of the mRNA spanning the polyadenylation site. Surprisingly two additional tags were found not corresponding to any known gene but showing close homology with the Tg SAGE tag sequence. We investigated whether these tags correspond to Tg mRNA molecules with alternative polyadenylation cleavage sites. Furthermore we studied if the preference of cleavage sites used in Tg mRNA follows the consensus for polyadenylation in eukaryotic mRNAs.

MATERIAL & METHODS

SAGE library

The library was constructed from human normal thyroid tissue according to original protocol (9) and as described previously (8).

3' RACE PCR

1st-strand RACE cDNA was synthesized from the same thyroid mRNA as used in the SAGE library. As a primer oligo-dT <sub>10-18</sub> with a 3' linker

<sup>5</sup>GCAcGcAgAgAttCttGGaTCC was used (10). The linker sequence was used as a reverse priming site in subsequent amplification of the Tg 3' region. Tg-specific forward primer Tg-2F <sup>5</sup>GAGAAgATCTCCTAAgCCTC was designed according to Tg cDNA sequence (GenBank accession no. U93033) generating a PCR product of about 200bp. PCR amplification was performed using standard conditions, 2mM MgCl<sub>2</sub> and 1'95°C, 1'55°C, 1'72°C, for 30 cycles.

Cloning of 3'RACE Tg-fragments

Tg 3'RACE PCR fragments were cloned using the pGEM/T-easy vector (Promega). Ligation reaction was transformed in DH5α cells and plated. DNA from Tg-positive clones was isolated using the Wizard Mini-Prep (Promega).

Amplification of Tg exon 48

Genomic DNA was isolated from the same thyroid tissue as used for the 3'RACE experiments using Trizol (Gibco-BRL) following manufacturer's protocol. 100 ng of DNA was used in a PCR amplification of exon 48 using intron-located primers with M13 linkers attached. Tg48-for: <sup>5</sup>AGAGAAgTCCtatCTGGCTTG, Tg48-rev: <sup>5</sup>CtGgTCATACAGATGCTCAT (GenBank accession no. AF080484).

Sequencing

Tg-RACE clones and exon 48 PCR product were sequenced with the Dyenamic Direct
cycle sequencing kit (Perkin Elmer) using M13 forward and reverse priming sites in
the vector or PCR fragment. Samples were run on an ABI377XL Automatic
Sequencer (Perkin Elmer) and analyzed using Sequence Analysis 3.0 software.

RESULTS & DISCUSSION

From the SAGE library results (8) the sequence tag corresponding to Thyroglobulin
(Tg) mRNA presented itself as $^5$CGGTGAAAAA and was scored 210 times out of a
total number of 10994 sequenced tags. Two tags not corresponding to any known
human cDNA, $^5$CGGTAAGCA and $^5$CGGAAAAAAA, were scored respectively 54
and 9 times. Because of the similarity of these tags to the 'wild-type' SAGE tag we
opted to clone Tg-3'RACE fragments to determine whether these tags corresponded
to Tg mRNA. In cloning 3'RACE fragments a representation of the Tg mRNA pool
can be found concerning the region of interest, namely the polyadenylation region.
Because priming sites are flanking this region no selection on sequence differences
will be made in the amplification step.

<table>
<thead>
<tr>
<th>SAGE tag</th>
<th>Freq</th>
<th>Sequence 3'RACE clone</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tg-tag1</td>
<td>210(77%)</td>
<td>$^5$CGGTGAAAAA</td>
<td>45 (73%)</td>
</tr>
<tr>
<td>Tg-tag2</td>
<td>54(20%)</td>
<td>$^5$CGGTAAGCA</td>
<td>9 (15%)</td>
</tr>
<tr>
<td>Tg-tag3</td>
<td>9(3%)</td>
<td>$^5$CGGAAAAAAA</td>
<td>3 (5%)</td>
</tr>
</tbody>
</table>

Figure 2 Alternative cleavage sites in thyroglobulin mRNA. Results from SAGE and 3'RACE are
summarized. Sequence for each variant is shown together with their relative abundance.
SAGE-tag frequency is out of a total of 10994 tags, 3'RACE-clone frequency is out of a total of 62 clones.

As is shown in figure 2, after sequencing 62 Tg 3'RACE-clones we found 4 different
polyadenylation sites. 45 correspond to the 'wild-type' Tg mRNA, so $^5$CGGTGAAAAA
(Tg-tag1), 9 clones have a polyadenylation site 4 bp downstream and corresponded
to $^5$CGGTAAGCA (Tg-tag2) and 5 clones show a polyadenylation site 3 bp
upstream, $^{5'}$CGGAAAAAAA (Tg-tag3). Additionally, 3 Tg cDNA-clones with a polyadenylation site 16bp downstream of the 'wild-type' cleavage site were found showing a fourth polyadenylation variant. Because in this case the first 10bp downstream from the Nlalll-restriction site (CATG) is similar to Tg-tag2 this variant could not be detected in the SAGE results. The extra 16bp Tg 3'UTR in these 3 clones correspond to the sequence as published for the last Tg coding exon (GenBank accession no. AF080484). The abundance of the four different Tg mRNA molecules is similar to the abundance of the respective tags in the SAGE expression library (fig. 2), also confirming the SAGE tags to arise from Tg mRNA. The 150bp sequence upstream of the polyadenylation site was identical for all 62 clones and to the Tg genomic sequence, indicating that there are no mutations in the upstream region that could be responsible for the variation in cleavage site selection. To exclude any mutations in the pre-mRNA sequence downstream from the polya-site putatively responsible for the variance in polyadenylation, we sequenced the last coding exon of Tg in the thyroid tissue used in these experiments. The genomic sequence of Tg in this particular thyroid gland is identical to the published Tg exon 48 sequence (results not shown), ruling out this possibility. After examination of the downstream pre-mRNA sequence present in the genomic sequence of the last Tg exon 48, a consensus pattern concerning the wild-type polyadenylation (4) could be found (fig. 3).

Figure 3 Genomic sequence (accession no. AF080484) of the last coding exon (48) of the human thyroglobulin gene. Consensus regulating elements are indicated as well as the 4 described cleavage sites. Smal font is coding sequence, large font represents pre-mRNA containing regulating polyadenylation signals.
The AAUAAA signal is present and for the three most abundant polyadenylation variants the distance to the polyA-site is within the defined consensus distance (11-24bp). As for the infrequent fourth variant, with 28bp the distance to the polyA-site may explain its low abundance. Following consensus downstream of the cleavage site a putative GU-rich site is present 18-23bp downstream of the Tg wild-type polyA-site, while another is present between 39-44bp from the wild-type polyA-site (fig. 3). It seems that the first putative GU-rich region 5'TGTGTG would be the most likely candidate responsible for the three most abundant cleavage sites, since the distance to this GU-rich region is 13-20bp. The fourth variant is only 6bp from this region and probably too close. It may use the second putative GU-rich region located 27bp from its cleavage site. This second GU-rich region may have a lower preference for the polyadenylation machinery since the expression of this last variant is lower compared to the three most abundant polyA-sites probably using the most upstream GU-rich region. The variation between the three most abundant polyadenylation variants possibly using the same upstream (AAUAAA) and downstream (UGUGUG) signal can be explained by the presence of several possible polyA-sites in the actual cleavage region. The first A seems to be the preferred site of cleavage since this variant has the highest expression level compared to all other variants.

In summary, it seems that the variation in polyadenylation/cleavage site in the human thyroglobulin mRNA is due to an in vivo variance caused by the presence of different downstream GU-rich regions such as in the case of the 16bp downstream variant. The three polyadenylation variants in the same cleavage region may be due to inaccuracy of the cleavage enzyme's cleaving at different nucleotides. This phenomenon may be tissue-, gene- or cleavage enzyme-specific and it would be interesting to study whether it influences further processing of the mRNA in the cell.