RNA (1997), *3*:697–701. Cambridge University Press. Printed in the USA.

LETTER TO EDITOR

Kissing loops hide premature termination codons in pre-mRNA of selenoprotein genes and in genes containing programmed ribosomal frameshifts

STEEN KNUDSEN and SØREN BRUNAK
Center for Biological Sequence Analysis, Department of Chemistry, The Technical University of Denmark,
DK-2800 Lyngby, Denmark

Keywords: mRNA stability; nuclear scanning; ribosomal frameshift; RNA secondary structure; selenoproteins; splicing

Premature termination codons (PTCs) have been reported recently to interfere with mRNA stability and splicing in the nucleus of eukaryotic cells (Maquat, 1995). To date, the mechanism by which such PTCs affect intron splicing and mRNA stability is still in dispute. Some experiments (Carter et al., 1996) suggest a nuclear scanning mechanism for PTCs, resulting in alternative or reduced splicing and/or increased mRNA decay when they are found. Other experiments, such as recent experiments in yeast (Zhang et al., 1997), suggest that recognition of PTCs takes place within the polysome compartment of the cell, and lead to cytoplasmic degradation of mRNA-containing PTCs.

The open reading frame of selenoprotein genes and genes containing programmed ribosomal frameshifts normally includes one or more PTCs. In selenoprotein genes, UGA triplets are well known for their role of encoding selenocysteine residues. An investigation of the genomic DNA of selenoprotein genes revealed that all known examples have introns. This means potentially that intron splicing and mRNA stability could be impaired.

Using free energy minimization, we discovered a general pre-mRNA secondary structure forming a "kissing loop" conformation, which incorporates part of the exon containing the PTC as well as part of the downstream intron sequence. The selenocysteine triplet UGA is incorporated in this novel double hairpin structure, which spans the exon/intron junction in all genes investigated. The RNA secondary structure is completely different from the 3′ UTR SECIS element (Berry et al., 1991; Walczak et al., 1996) required for translation of selenocysteine codons.

It is remarkable that a similar kissing loop structure was found in a gene containing a programmed ribosomal frameshift. A premature UGA codon that terminates the unrestored reading frame of the rat and mouse *oaz* genes is also incorporated in this type of RNA secondary structure.

The finding may explain how selenoproteins and proteins originating from translation anomalies can be produced in eukaryotes from genes containing intervening sequences. Together the results offer new insight into the mechanism of nuclear scanning for nonsense codons (Dietz & Kendzior, 1994; Carter et al., 1996), as well as new insight into the differences between nuclear scanning and ribosomal translation. Overall, the results are consistent with the nuclear scanning model.

We extracted the genomic DNA of all intron-containing genes with "natural" PTCs annotated in GenBank. This class did not include retroviral frameshifts and frameshifts in eukaryotic transposable element transcripts, because these genes were found not to contain any introns.

A study of all nine characterized vertebrate selenoprotein genes deposited as genomic DNA revealed that they all had introns, and that the UGA codons were located at preferred positions in the exons. In 8 of 12 cases, the UGA codons were found to appear within a narrow distance interval between 80 and 125 bp from the downstream 5′ splice site. The distance distribution showed a peak at 108 nt. Further analysis of the sequence features led to the discovery of a conserved signal in the 5′ end of the introns, as well as a signal in

Reprint requests to: Søren Brunak, Center for Biological Sequence Analysis, Department of Chemistry, Building 206, The Technical University of Denmark, DK-2800, Lyngby, Denmark; e-mail: brunak@cbs.dtu.dk.

the exons extending all the way up to the selenocysteine codon (Fig. 1).

Using a standard hidden Markov model (Eddy, 1996; Hughey & Krogh, 1996) built by extracting the information available in all nine genes, we scanned GenBank for additional entries containing the sequence pattern. A hidden Markov model contains information about the probabilities of individual nucleotides, as well as insertions and deletions, at given positions of a model of a specific signal. After building such a model, a multiple alignment can be produced by aligning the sequences that were used to generate the model to it. Comparing GenBank sequences to the model gives the log-likelihood that the sequence was produced by the model, i.e., shares properties with the sequences used to build the model. With this approach, we found 11 5′-expressed sequence tags, of which most did not match any known DNA or protein sequence (Table 1).

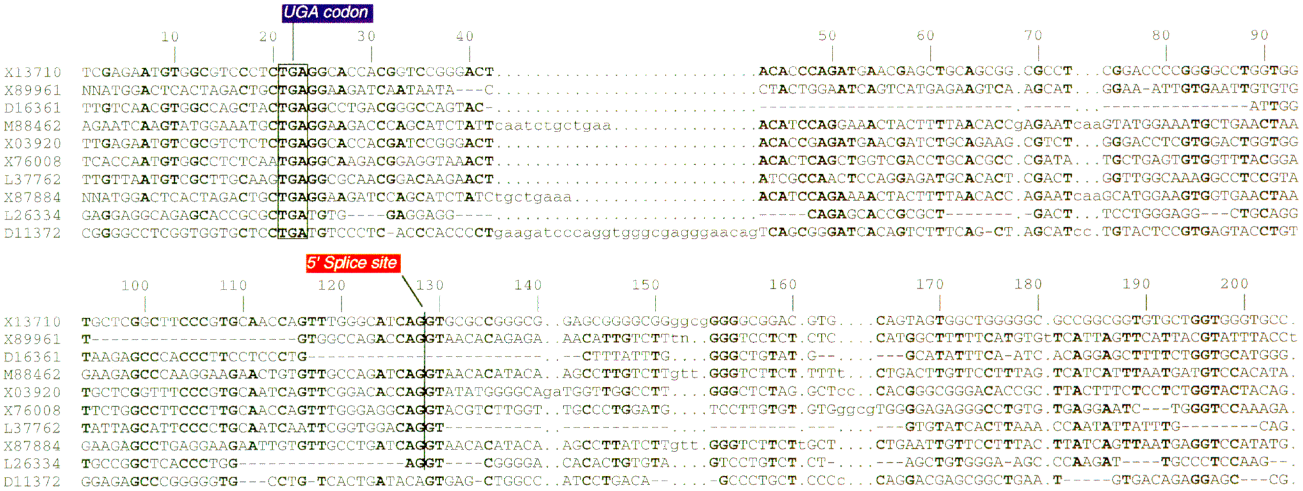
The predicted local RNA secondary structure of the sequences showed some remarkable conserved features. A second hidden Markov model was used to align the predicted secondary structures of 20 sequences (selenoprotein sequences and sequences found by the GenBank search; see Table 1). The structural alignment showed five conserved loops. The common secondary structure with the lowest free energy was one that put the selenocysteine codon in one hairpin structure and the 5′ splice site in another. Two examples are shown in Figure 2. Among the nine known selenoprotein genes, two contained more than one selenocysteine codon. It was remarkable that, in both cases, they occurred in the same exon and were embedded in adjacent hairpins. In a mouse selenoprotein (MUSSEL) that has three selenocysteines in one exon, the UGA codons appeared in separate hairpins, where the last hairpin contained the 5′ splice site. Downstream from the two conserved hairpins, we regularly found a conserved run of pyrimidines followed by a short stem with the sequence GU in the loop (Fig. 2).

Another exceptional feature of the individual secondary structures is the potential for perfect base pairing between the two hairpins (Fig. 2). A systematic search showed that all nine nonhomologous sequences had this striking base complementarity between the loops. Such base pairing between loops in separate secondary structure elements has been predicted previously in 16S ribosomal RNA (Gutell, 1993).

Numerous experiments have demonstrated that nonsense codons in the last exon have no effect on mRNA stability (Maquat, 1995). As expected, we found none of the primary and secondary structure features described here associated with a putative selenocysteine codon in the last exon of the *Drosophila* OAF protein pre-mRNA (L31349).

Several of the ESTs aligned with the splice sites of the genomic sequences and matched the intronic part of the novel signal. That indicates the presence of alternative splicing in these ESTs, as was found in a rat selenoprotein cDNA that was isolated in several different versions (Karimpour et al., 1992).

If selenocysteine codons are hidden from the nuclear scanning apparatus in the nucleus, then what about PTCs that appear in the unrestored reading frame after a programmed ribosomal frameshift? Either the nuclear scanning apparatus has to recognize the frameshift signal and resume scanning in the new frame, or



```
                10        20         30        40                                50        60         70           80        90
X13710  TCGAGAATGTGGCGTCCCTCTGAGGCACCACGGTCCGGGACT...........................ACACCCAGATGAACGAGCTGCAGCGG.CGCCT...CGGACCCCGGGGCCTGGTGG
X89961  NNATGGACTCACTAGACTGCTGAGGAAGATCAATAATA---C...........................CTACTGGAATCAGTCATGAGAAGTCA.AGCAT...GGAA-ATTGTGAATTGTGTG
D16361  TTGTCAACGTGGCCAGCTACTGAGGCCTGACGGGCCAGTAC-...........................--------------------------.-----...--------------ATTGG
M88462  AGAATCAAGTATGGAAATGCTGAGGAAGACCCAGCATCTATTcaatctgctgaa...............ACATCCAGGAAACTACTTTTAACACCgAGAATcaaGTATGGAAATGCTGAACTAA
X03920  TTGAGAATGTCGCGTCTCTCTGAGGCACCACGATCCGGGACT...........................ACACCGAGATGAACGATCTGCAGAAG.CGTCT...GGGACCTCGTGGACTGGTGG
X76008  TCACCAATGTGGCCTCTCAATGAGGCAAGACGGAGGTAAACT...........................ACACTCAGCTGGTCGACCTGCACGCC.CGATA...TGCTGAGTGTGGTTTACGGA
L37762  TTGTTAATGTCGCTTGCAAGTGAGGCGCAACGGACAAGAACT...........................ATCGCCAACTCCAGGAGATGCACACT.CGACT...GGTTGGCAAAGGCCTCCGTA
X87884  NNATGGACTCACTAGACTGCTGAGGAAGATCCAGCATCTATCtgctgaaa.................ACATCCAGAAAACTACTTTTAACACC.AGAATcaaGCATGGAAGTGGTGAACTAA
L26334  GAGGAGGCAGAGCACCGCGCTGATGTG----GAGGAGG----...........................-----CAGAGCACCGCGCT-------.-GACT...TCCTGGGAGG---CTGCAGG
D11372  CGGGGCCTCGGTGGTGCTCCTGATGTCCCTC-ACCCACCCCTgaagatcccaggtgggcgagggaacagTCAGCGGGATCACAGTCTTTCAG-CT.AGCATcc.TGTACTCCGTGAGTACCTGT

            100       110       120       130       140          150          160           170       180         190       200
X13710  TGCTCGGCTTCCCGTGCAACCAGTTTGGGCATCAGGTGCGCCGGGCG..GAGCGGGGCGGggcgGGGGCGGAC.GTG....CAGTAGTGGCTGGGGGC.GCCGGCGGTGTGCTGGTGGGTGCC.
X89961  T---------------------GTGGCCAGACCAGGTAACACAGAGA..AACATTGTCTTtn..GGGTCCTCT.CTC....CATGGCTTTTTCATGTGtTCATTAGTTCATTACGTATTTACCt
D16361  TAAGAGCCCACCCTTCCTCCCTG------------------------..--CTTTATTTG....GGGCTGTAT.G--....---GCATATTTCA-ATC.ACAGGAGCTTTTCTGGTGCATGGG.
M88462  GAAGAGCCCAAGGAAGAACTGTGTTGCCAGATCAGGTAACACATACA..AGCCTTGTCTTgtt.GGGTCTTCT.TTTt...CTGACTTGTTCCTTTAG.TCATCATTTAATGATGTCCACATA.
X03920  TGCTCGGTTTCCCGTGCAATCAGTTCGGACACCAGGTATATGGGGCAgaTGGTTGGCCTT....GGGCTCTAG.GCTcc..CACGGGCGGGACACCGC.TTACTTTCTCCTCTGGTACTACAG.
X76008  TTCTGGCCTTCCCTTGCAACCAGTTTGGGAGGCAGGTACGTCTTGGT..TGCCCTGGATG....TCCTTGTGT.GTGggcgTGGGGAGAGGGCCTGTG.TGAGGAATC---TGGGTCCAAAGA.
L37762  TATTAGCATTCCCCTGCAATCAATTCGGTGGACAGGT----------..-----------....---------.---....---GTGTATCACTTAAA.CCAATATTATTTG--------CAG.
X87884  GAAGAGCCTGAGGAAGAATTGTGTTGCCTGATCAGGTAACACATACA..AGCCTTATCTTgtt.GGGTCTTCTtGCT....CTGAATTGTTCCTTTAC.TTATCAGTTAATGAGGTCCATATG.
L26334  TGCCGGCTCACCCTGG-----------------AGGT----CGGGGA..CACACTGTGTA....GTCCTGTCT.CT-....---AGCTGTGGGA-AGC.CCAAGAT----TGCCCTCCAAG--.
D11372  GGAGAGCCCGGGGGTG---CCTG-TCACTGATACAGTGAG-CTGGCC..ATCCTGACA--....-GCCCTGCT.CCCc...CAGGACGAGCGGCTGAA.T-----GTGACAGAGGAGC---CG.
```

FIGURE 1. Multiple sequence alignment of the region surrounding the TGA codon in GenBank selenoprotein genes and the rat *oaz* sequence in which the kissing loop signal also was found. Conserved nucleotides are emphasized in bold-face when at least 6 of 10 are identical. The selenocysteine codon TGA is also emphasized in bold, and the splice site is indicated by a vertical line. Pairwise treealign (Hein, 1990) and fasta (Pearson & Lipman, 1988) alignments of all nine selenoproteins and their DNA showed very limited sequence identity overall (only 2 pairwise alignments of 81 had significant sequence identity), ruling out the possibility that the observed local sequence conservation was merely a reflection of common functional identity or evolutionary history.

TABLE 1. List of sequences either annotated as selenoproteins in GenBank or found by scanning GenBank with a hidden Markov model comprising 200 main states (Eddy, 1996; Hughey & Krogh, 1996).[a]

Name	Acc #	Location of TGA	Description
1. HSPEROXP	X13710	107 bp from first 5′ ss	Human glutathione peroxidase gene
2. HSMCSGEN2	X89961	83 bp from first 5′ ss	Human mitochondrial capsule selenoprotein gene(*)
3. HUMGPXP2	D16361	25 bp from second 5′ ss	Human plasma glutathione peroxidase gene
4. MUSSEL	M88462	124, 94, and 43 bp	Mouse seleno-protein gene (MCS) from first 5′ ss
5. MMGSHPX	X03920	109 bp from first 5′ ss	Mouse glutathione peroxidase gene
6. SSPHGPX	X76008	108 bp from third 5′ ss	*S. scrofa* phospholipid hydroperoxide glutathione peroxidase gene
7. SCMGPX1A	L37762	108 bp from third 5′ ss	*Schistosoma mansoni* glutathione peroxidase gene
8. RNMCSGENE	X87884	119 bp from first 5′ ss	Rat mitochondrial capsule selenoprotein gene(*)
9. GPIZFP	L26334	48, 21 bp from 5′ ss	*Cavia porcellus* zinc finger protein gene
10. HUMGLPEX	M83094	106 from first 5′ ss	Human cytosolic selenium-dependent glutathione peroxidase gene
11. RATODCAC	D11372	85 bp from second 5′ ss	Rat ornithine decarboxylase antizyme
12. MMU84291	U84291	85 bp from second 5′ ss	Mouse ornithine decarboxylase antizyme
13. EST	H70052	123 bp from put. 5′ ss	Human 5′ EST from mRNA
14. EST	H88181	84 bp from put. 5′ ss	Human 5′ EST from mRNA
15. EST	H61655	—	Human 5′ EST from mRNA
16. EST	H99005	—	Human 5′ EST from mRNA
17. EST	H33776	97 bp from put. 5′ ss	Human 5′ EST from mRNA
18. EST	R21003	83 bp from put. 5′ ss	Human 5′ EST from mRNA
19. EST	R47070	91 bp from put 5′ ss	Rat 5′ EST. cDNA from incisor tissue
20. EST	R65113	—	Human 5′ EST from mRNA
21. EST	R69653	—	Human 5′ EST from mRNA
22. EST	T54820	102 bp from put. 5′ ss	Human 5′ EST from mRNA
23. EST	N38469	—	*A. thaliana* 5′ EST from mRNA
24. 16S PCR	X84460	—	PCR amplification from unkn. organism using probes for 16S rRNA

[a] This Markov model was run against GenBank (rel. 94). Sequences with scores significantly higher than average were investigated further by alignment to the known selenoprotein genes and by searching for similarities to genes deposited in GenBank. Sequences 1–9 were used for HMM training and are shown in the alignment in Figure 1, together with an example of a frameshifting gene, sequence 11. Sequence 10, a homologue of sequence 5, appeared in a later release of GenBank. Sequences 13–24 were found in a GenBank search with the trained HMM. All sequences have the conserved secondary structure. The secondary structures of sequences 2, 7, and 11 are shown in Figure 2A, B, and C, respectively. The distance between the TGA codon and the 5′ splice site, if annotated or predicted, is given.

the codons that terminate the original frame have to be hidden from nuclear scanning. The search in GenBank for genes of this class revealed that both known examples—the genes *oaz* in rat and mouse—had introns. In these genes, a PTC appears between a frameshifting signal and a 5′ splice site. We studied the pre-mRNA secondary structure of both rat and mouse *oaz* and found a structure very similar to that found in selenoprotein genes, suggesting that the frameshifting signal poses a problem to the nuclear scanning mechanism just as the selenocysteine codons may do (Fig. 2). Rat *oaz* pre-mRNA is included in the multiple alignment shown in Figure 1.

The secondary structure shown in Figure 2 may well be providing the basis for circumventing the nuclear scanning of selenocysteine encoding exons, thus presenting new independent evidence for the nuclear scanning mechanism. It is possible that the secondary structure is stabilized by protein binding, and/or that its function is mediated by interaction with proteins. As an example of the latter, the RNA SECIS element required for translational insertion of selenocysteine has been shown to interact with a protein (Hubert et al., 1996; Lesoon et al., 1997).

Excluding the loop-to-loop interaction, the local free energies of the three structures shown in Figure 2 are: −15 kcal/mol (Fig. 2A); −22 kcal/mol (Fig. 2B); −58 kcal/mol (Fig. 2C). For comparison, a 3′ UTR SECIS element of comparable length had a free energy of −44 kcal/mol (Berry et al., 1991). Even if it is unclear to what extent nuclear scanning complexes may share components with the cytoplasmic translation machinery, it appears unlikely that these structures by themselves are stable enough to directly prevent denaturation by a codon-scanning mechanism.

Another example of a RNA–protein interaction which may prove to be relevant in this splicing associated context comes from the U5 snRNP, where the Sm protein-binding site consists of a short run of Us, followed by a short hairpin (Jones & Guthrie, 1990). Following the two hairpins shown in Figure 2, we also consistently found a run of pyrimidines followed by a short stem with the conserved sequence GU in the loop.

The kissing loop structure described in this paper falls under the structural class of pseudoknots. An interesting example of protein binding to such structures has been described in *Escherichia coli* (Baker & Draper, 1995). Protein S4 blocks the translation of the alpha operon by binding to an mRNA pseudoknot. Protein S4 in *E. coli* binds not only to a pseudoknot in the mRNA, but also to 16S rRNA. Notably, several 16S

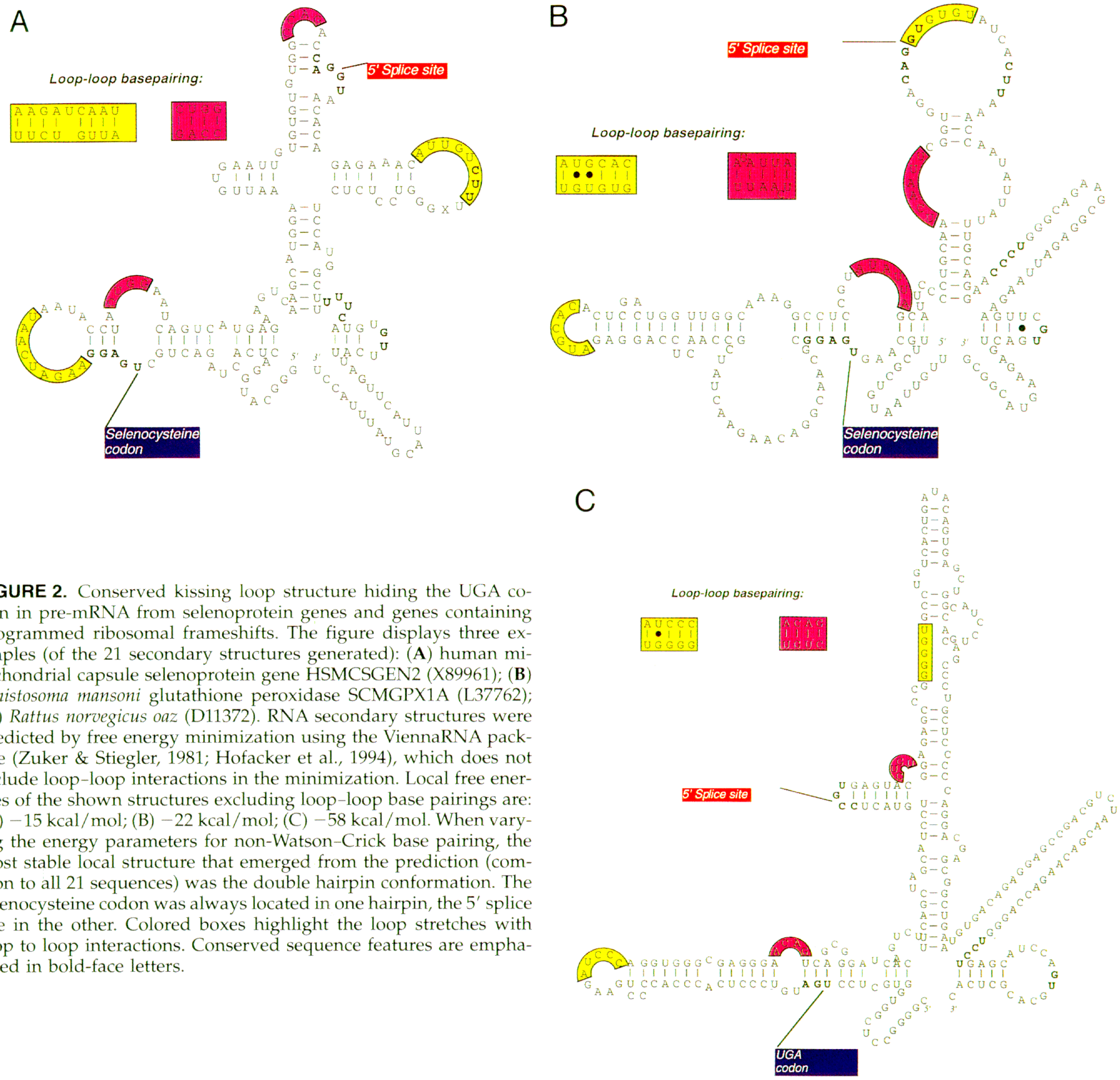


FIGURE 2. Conserved kissing loop structure hiding the UGA codon in pre-mRNA from selenoprotein genes and genes containing programmed ribosomal frameshifts. The figure displays three examples (of the 21 secondary structures generated): (**A**) human mitochondrial capsule selenoprotein gene HSMCSGEN2 (X89961); (**B**) *Schistosoma mansoni* glutathione peroxidase SCMGPX1A (L37762); (**C**) *Rattus norvegicus oaz* (D11372). RNA secondary structures were predicted by free energy minimization using the ViennaRNA package (Zuker & Stiegler, 1981; Hofacker et al., 1994), which does not include loop–loop interactions in the minimization. Local free energies of the shown structures excluding loop–loop base pairings are: (A) −15 kcal/mol; (B) −22 kcal/mol; (C) −58 kcal/mol. When varying the energy parameters for non-Watson–Crick base pairing, the most stable local structure that emerged from the prediction (common to all 21 sequences) was the double hairpin conformation. The selenocysteine codon was always located in one hairpin, the 5′ splice site in the other. Colored boxes highlight the loop stretches with loop to loop interactions. Conserved sequence features are emphasized in bold-face letters.

rRNA-like sequences from GenBank aligned well to the selenoproteins. Consequently, we also aligned *E. coli* 16S rRNA and human 18S rRNA. Although they do not have a high score (log-likelihood) from the hidden Markov model, they actually align at the selenocysteine codon even though they are not known to encode any proteins (data not shown). Further investigation of the molecular basis of nuclear scanning could possibly reveal whether an S4-like protein interacts with the kissing loop structure in selenoprotein pre-mRNA.

The fact that the conserved kissing loop structure involves noncoding intron sequence supports the notion that nonsense codons are detected as a nuclear event (Li et al., 1997). If nuclear scanning shares components with the cytoplasmic translation machinery, however, cytoplasmic translation must remain uninhibited. Indeed, splicing removes the second hairpin of the conserved structure, thereby removing the possible obstacle to translation. This holds true for both selenoprotein genes and the intron-containing genes with programmed ribosomal frameshifts.

The observation that the conserved kissing loop structure is removed by splicing before export to the cytoplasm and subsequent translation suggests also that it may interfere with cytoplasmic translation. That would not be surprising if nuclear scanning proceeded with components that are shared with the ribosome.

Messenger RNA stability experiments have showed that the detrimental effect of PTCs is abolished upon addition of the translation inhibitor cycloheximide. This observation has been taken as support for PTC recognition during translation in the cytoplasm. An alternative interpretation, consistent with our results, is that nuclear scanning components are blocked by cycloheximide as well.

Indeed, the 40S ribosomal subunit is renown for its ability to carry out scanning for a start codon on mRNA (Kozak, 1989). The small subunit has also been proposed to be the main frame keeping agent (Trifonov, 1987). Although 80S ribosomes are known to resolve any secondary structure encountered during translation, the 40S ribosomal subunit has been shown to be able to bypass a start codon sequestered by a stem-loop structure in the cauliflower mosaic virus 35S RNA (Futterer et al., 1993). The principle behind this shunting mechanism is very similar to the PTC circumvention scenario described above.

ACKNOWLEDGMENTS

This work was supported by a grant from the Danish National Research Foundation. We thank Kristoffeer Rapacki for assistance in GenBank extraction.

REFERENCES

Baker A, Draper D. 1995. Messenger RNA recognition by fragments of ribosomal protein S4. *J Biol Chem 270*:22939–22945.

Berry M, Banu L, Chen Y, Mandel S, Kieffer J, Harvey J, Larsen P. 1991. Recognition of UGA as a selenocysteine codon in type I deiodinase requires sequences in the 3′ untranslated region. *Nature 353*:273–276.

Carter M, Shulin L, Wilkinson M. 1996. A splicing-dependent regulatory mechanism that detects translation signals. *EMBO J 15*:5965–5975.

Dietz H, Kendzior R. 1994. Maintenance of an open reading frame as an additional level of scrutiny during splice site selection. *Nature Gen 8*:183–188.

Eddy S. 1996. Hidden Markov models. *Curr Opin Struct Biol 6*:361–365.

Futterer J, Kiss-Laszlo Z, Hohn T. 1993. Nonlinear ribosome migration on cauliflower mosaic virus 35S RNA. *Cell 73*:789–802.

Gutell R. 1993. Collection of small subunit (16S- and 16S-like) ribosomal RNA structures. *Nucleic Acids Res 21*:3051–3054.

Hein J. 1990. A unified approach to alignments and phylogenies. *Methods Enzymol 183*:625–645.

Hofacker I, Fontana W, Stadler P, Bonherffer S, Tacker M, Schuster P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte f Chemie 125*:167–188.

Hubert N, Walczak R, Carbon P, Krol A. 1996. A protein binds the selenocysteine insertion element in the 3′-UTR of mammalian selenoprotein mRNAs. *Nucleic Acids Res 24*:464–469.

Hughey R, Krogh A. 1996. Hidden Markov models for sequence analysis: Extension of the basic method. *CABIOS 12*:95–107.

Jones M, Guthrie C. 1990. Unexpected flexibility in an evolutionarily conserved protein–RNA interaction: Genetic analysis of the Sm binding site. *EMBO J 9*:2555–2561.

Karimpour I, Cutler M, Shih D, Smith J, Kleene K. 1992. Sequence of the gene encoding the mitochondrial capsule selenoprotein of mouse sperm: Identification of three in-phase TGA selenocysteine codons. *DNA Cell Biol 9*:693–699.

Kozak M. 1989. The scanning model for translation: An update. *J Cell Biol 108*:229–241.

Lesoon A, Mehta A, Singh R, Chisolm G, Driscoll D. 1997. An RNA-binding protein recognizes a mammalian selenocysteine insertion sequence element required for cotranslational incorporation of selenocysteine. *Mol Cell Biol 17*:1977–1985.

Li S, Leonard D, Wilkinson M. 1997. T cell receptor (TCR) mini-gene mRNA expression regulated by nonsense codons: A nuclear-associated translation-like mechanism. *J Exp Med 185*:985–992.

Maquat L. 1995. When cells stop making sense: Effect of nonsense codons on mRNA metabolism in vertebrate cells. *RNA 1*:453–465.

Pearson W, Lipman D. 1988. Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA 85*:2444–2448.

Trifonov E. 1987. Translation framing code and frame-monitoring mechanism as suggested by the analysis of mRNA and 16S rRNA nucleotide sequences. *J Mol Biol 194*:643–652.

Walczak R, Westhof E, Carbon P, Krol A. 1996. A novel structural motif in the selenocysteine insertion element of eukaryotic selenoprotein mRNAs. *RNA 2*:367–379.

Zhang S, Welch E, Hagan K, Brown A, Peltz S, Jacobson A. 1997. Polysome-associated mRNAs are substrates for the nonsense-mediated mRNA decay pathway in *Saccharomyces cerevisiae*. *RNA 3*:234–244.

Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamic and auxiliary information. *Nucleic Acids Res 9*:133–148.