RNA editing of the transcript coding for subunit 4 of NADH dehydrogenase in wheat mitochondria: uneven distribution of the editing sites among the four exons

Lorenzo Lamattina and Jean Michel Grienenberger*
Institut de Biologie Moléculaire des Plantes du CNRS, Université Louis Pasteur, 12 rue du Général Zimmer, 67084 Strasbourg Cedex, France

Received March 26, 1991; Revised and Accepted May 15, 1991 EMBL accession nos X57163 and X57164

ABSTRACT

The wheat mitochondrial (mt) NADH dehydrogenase subunit 4 gene (*nad4*) has been localized and sequenced. This gene, about 8 kb long, is composed of four exons separated by three class II introns. The *nad4* gene exists as a single copy in the wheat mitochondrial genome and it is transcribed into one abundant mRNA of 1.8 kb, whose extremities have been mapped. The complete cDNA sequence corresponding to the *nad4* transcript has been determined by combining the direct sequencing of uncloned cDNA and a method involving cDNA synthesis and PCR amplification using specific oligonucleotides as primers, followed by cloning and sequencing of the amplification product. Comparison of the genomic sequence with that of the cDNA shows that all *nad4* transcripts are fully edited at 23 positions, with an uneven distribution of the editing sites between the different exons: While exon 1 and exon 4 are extensively edited (with a change of 11% of the amino acid sequence), exon 2 is not edited at all and exon 3 is 0.5% edited. This uneven distribution is discussed.

INTRODUCTION

By immunological studies, seven open reading frames of the human mt genome have been identified as the genes coding for seven polypeptides which are part of complex I, the NADH dehydrogenase (1).These genes have been found in nearly all mitochondrial genomes except in yeast (2). In higher plant mitochondria, the genes coding for the subunits 1, 2, 3 and 5 of NADH dehydrogenase have been recently identified (3–7). The partial sequence of a putative gene coding for the subunit 4 (*nad4*) has also been reported in soybean mitochondria (8). By comparison of the deduced polypeptide sequence with known protein sequences, it was shown that this partial sequence codes for the C-terminal part of *nad4* and also that *nad4* is interrupted. Using the soybean sequence as a probe, a similar sequence has been identified in wheat mitochondria (9) and the presence of a 3398 nt long class II intron has been shown.

Recently, we (9,10) and others (11,12) have shown that RNA editing, a newly discovered way of modulating genetic information, is occurring in plant mitochondria. This activity results into the presence of a U in the mRNA at specific locations where a C is encoded by the DNA. These conversions result in an mRNA whose nucleotide sequence is different from that of the gene and it is therefore not possible to deduce the protein sequence directly from the gene sequence. The determination of the mRNA sequence, usually through cDNA sequencing, is now a necessary step to obtain the correct aminoacid sequence of a protein encoded by the plant mt genome (10,11).

This work describes the study of a region of the wheat mitochondrial genome which contains the gene coding for subunit 4 of NADH dehydrogenase. This gene, which is about 8kb long, is composed of four exons separated by three class II introns. The complete sequence of this gene can be found in GenEMBL data bank under accession number X57164. In this report, we present the physical structure of this gene, the nucleotide sequence of the *nad4* cDNA with the complete characterization of the edited sites of the *nad4* transcript.

MATERIAL AND METHODS

Wheat mitochondrial DNA

The complete collection of wheat mt DNA *Sal* I clones was obtained from F. Quétier and B. Lejeune (Laboratoire de Biologie Moléculaire Végétale, Université Paris XI, Orsay, France).

Extraction and analysis of RNA

Wheat seeds (*Triticum aestivum*, variety Capitole) were grown in the dark at 25°C for 10 days. Total mt RNAs of the etiolated seedlings were extracted as described by Stern and Newton (13). RNAs were analysed by electrophoresis on formaldehyde-agarose gel and blotted onto nitrocellulose filters. Oligonucleotides probes were labelled with T4 polynucleotide kinase using (γ-^{32}P) ATP as described (14). Probes were hybridized to filters at 42°C in 50% formamide, 5×SSPE, 5×Denhardt's, 0.5% SDS, at least during 48 h. Washings were performed at high stringency (45°C, 0.1×SSPE and 0.1% SDS).

* To whom correspondence should be addressed

DNA cloning and sequencing

Subclones of wheat mt DNA were obtained in M13 vectors by transformation of *Escherichia coli* NM522 host strains. Standard recombinant techniques were followed (14). Sequencing of single-strand DNA was performed using the dideoxynucleotide chain termination method (15). Direct sequencing of PCR amplification products was carried out as described (15) with some modifications. The hybridization reaction contained about 10 ng of double-stranded PCR product purified through NACS columns (BRL), 50 mM Tris-HCl pH 7.5, 20 mM $MgCl_2$, 70 mM NaCl and 5 ng of ^{32}P-labelled primer in 10 μl . The reaction mixture was boiled for 10 min, immediately put at −70°C for 15 sec and rapidly mixed with 5.5 μl of the labelling solution containing 0.1 μM of each dNTP, 20 μM DTT and 3 U of T7 DNA polymerase. After 30 sec at room temperature, termination mix were added as in a conventional sequencing reaction (15) and the incubation was carried on at 37°C for another 2 min. The reaction was stopped by adding formamide solution and the products were analysed on a 6% polyacrylamide gel.

Direct sequence of uncloned cDNA

For cDNA sequencing, 60 μg of total wheat mt RNAs were hybridized with 4 ng of 5'-end ^{32}P-labelled specific oligonucleotides in 13 μl of a solution containing 250 mM KCl and 10 mM Tris-HCl pH 8.3. The solution was heated at 80°C during 3 min, followed by 60 min at 5°C below the denaturation temperature calculated for each oligonucleotide. The dideoxy chain termination method (15) was carried out using AMV reverse transcriptase in four tubes with 10 μl of a reaction mixture containing: 24 mM Tris-HCl pH 8.3, 16 mM $MgCl_2$, 8 mM DTT, 0.8 mM of each dNTP, 100 μg/ml actinomycin D, 3 μl of the primer/RNA solution and 1 μl of 1 mM of one of the ddNTP. The mixture was incubated at 45°C for 45 min, denaturated and the products were resolved on a 6% polyacrylamide gel.

S1 nuclease and primer extension mapping

For S1 nuclease mapping of 3'-end of E3, the 383 nt *Hpa*II-*Hae*III fragment was purified from agarose gel after digestion of 1.9 kb *Sac*I fragment (Figure 1a) and labelled with (α-^{32}P) dCTP (400Ci/ mmol, Amersham). The labelled coding strand was isolated in a denaturating polyacrylamide gel (14). After elution, the sample was precipitated with ethanol, dried and resuspended in 10 μl of 10 mM Tris pH 7.5, 1 mM EDTA. For S1 nuclease digestion, 40 μg of total mt RNAs were hybridized with about 20 ng of the purified single-stranded DNA in 10 mM Pipes pH 6.4, 0.4 M NaCl and 1 mM EDTA. After 3 h at 50°C, hybrids were incubated with 4000 U/ml or 2000 U/ml of S1 nuclease (Appligene) at 37°C for 60 min. The solution was phenol-extracted and ethanol precipitated. The products were analysed on a 6% polyacrylamide gel.

For S1 nuclease analysis of the 3'-end of *nad*4 transcript, the oligonucleotide O_9 (Figure 1a, Table I) was labelled at its 3'-end with (α-^{32}P)dCTP and Terminal Deoxynucleotidyl Transferase (TdT). The reaction mix contained 0.1 M potassium cacodylate (pH 7.2), 2 mM $CoCl_2$, 0.2 mM DTT, 3 mM (α-^{32}P) dCTP (3000 Ci/mmol, Amersham), 200 ng of oligonucleotide O_9 (47 mer) and 15 U of TdT in 10 μl of reaction volume. The reaction was carried out at 37°C for 2 min and immediately heated at 70°C to inactivate the enzyme. The products were separated on a 6% polyacrylamide gel in order to isolate the DNA species which have incorporated only one nucleotide. The 48 mer

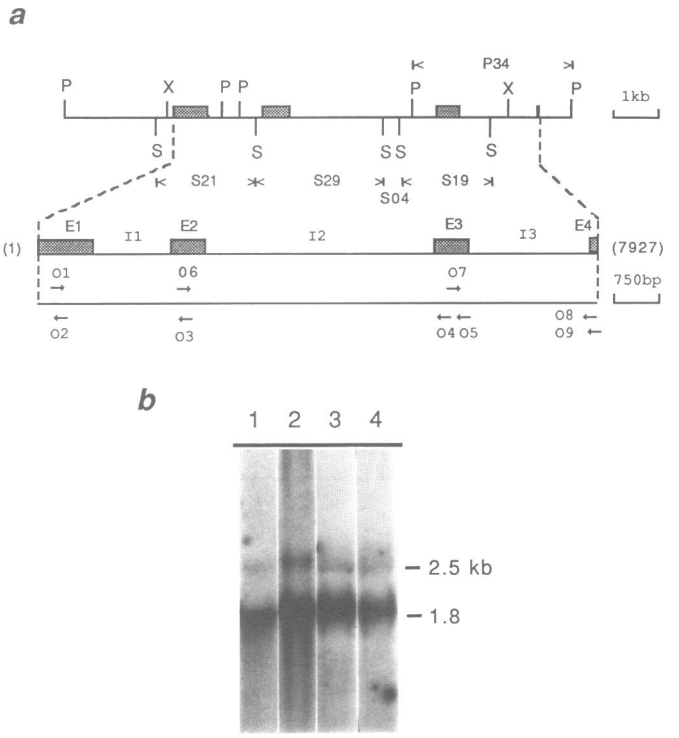

Figure 1. Organization of wheat mitochondrial *nad*4 gene and Northern analysis. **a**, partial restriction map of the 23 kb *Sal* I fragment (C3) of wheat mt DNA containing the *nad*4 gene. E1, exon 1; E2,exon 2; E3,exon 3 and E4, exon 4; I1, intron 1; I2, intron 2 and I3, intron 3. O_{1-9}, synthetic oligonucleotides used for DNA and RNA sequencing, for polymerase chain reaction and for Northern analysis. Relative locations of the primers are shown with the direction of chain elongation indicated by arrows. Numbers into parentheses indicate the coordinates of wheat mt genomic sequence, starting from 5'-end of exon 1 of *nad*4. S21, S29, S19,S04 and P34 indicate the fragments whose sequence has been determined . P= *Pst*I; S= *Sac*I; X= *Xho*I. **b**, Northern analysis of the wheat mt *nad*4 transcripts. About 50 μg of total wheat mt RNAs were probed with oligonucleotides internal to the four *nad*4 exons, O_2 (lane1); O_3 (lane 2); O_4 (lane 3) and O_8 (lane 4). Probe preparation is described in Methods. The sizes of the transcripts were determined using RNA size markers (BRL).

oligonucleotide was eluted from the gel, ethanol-precipitated and its length checked by comparison with the standard sequence of M13mp18. About 5 ng of 3'-end labelled and purified oligonucleotide were hybridized with 40 μg of total wheat mt RNAs in 10 mM Pipes pH 6.4 ; 0.4 M NaCl and 1 mM EDTA. The denaturation step was at 97°C for 3 min and the hybridization step was at 65°C for 60 min. After digestion with 150 U of S1 nuclease at 37°C for 30 min, the protected oligonucleotide was analysed on a 6% polyacryamide gel.

Primer extension analysis were performed using AMV reverse transcriptase as decribed above for the sequencing of uncloned cDNA without adding ddNTP in the reaction. After denaturation for 5 min at 95°C in 40% formamide, the reaction products were separated on a 6% polyacrylamide gel.

cDNA synthesis and PCR amplification

The PCR technique was used to amplify the cDNA corresponding to exons 1, 2 and 3 of *nad*4 transcript. Total mt RNAs (6 mg/ml) was treated with 100 U/ml of RNAse-free DNAse, during 45 min at 37°C. After phenol extraction and ethanol precipitation, the pellet was dried and 40 μg of mt RNAs were mixed with 150 ng of oligonucleotide O_5 (Figure 1a, Table I). Hybridization

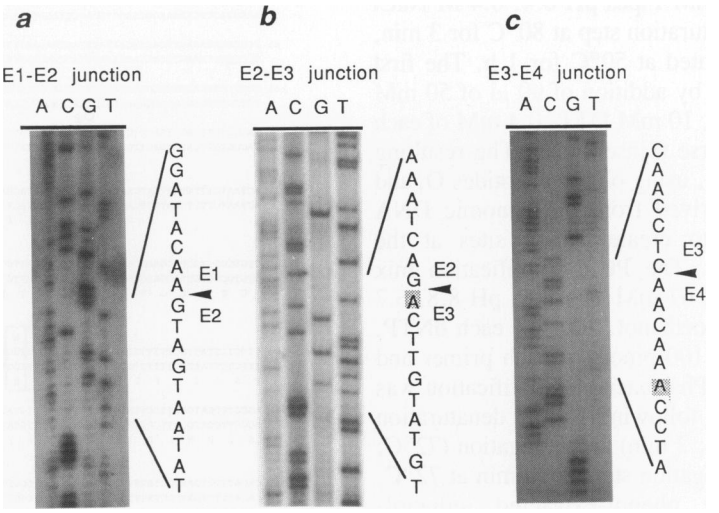

Figure 2. Direct, uncloned cDNA sequences of wheat mt *nad*4 transcripts at the junction sites of exons. The sequences shown correspond to the coding strand, complementary to the RNA. **a**, exon 1-exon 2 ; **b**, exon 2-exon 3 and **c**, exon 3-exon 4. For the cDNA sequences of a, b and c, O_3, O_4 and O_8 primers were respectively used. Arrows indicate the junction sites of exons. The shadowed As (corresponding to a U in the mRNA) indicate the sites which are edited and correspond to a C in the genomic DNA (Figure 4).

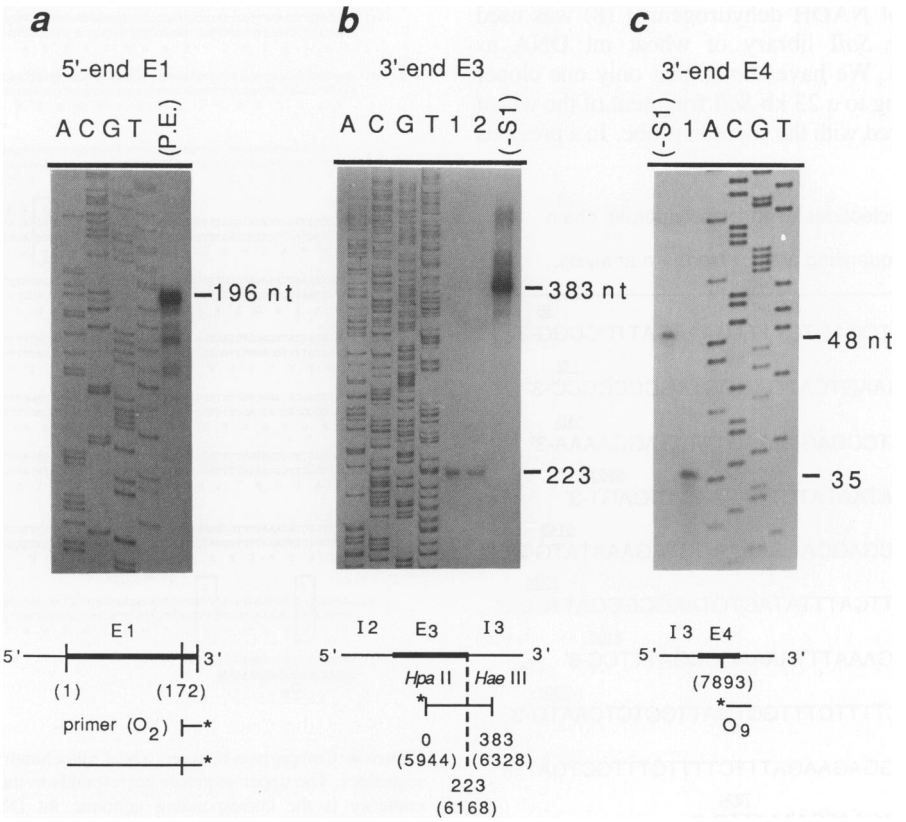

Figure 3. Primer extension analysis and S1 mapping of the wheat mt *nad*4 transcripts. **a**, 5′-end labelled O_2 was hybridized to total wheat mt RNAs. Primer extension by AMV reverse transcriptase was carried out as described in Materials and Methods. P.E., primer extension reaction. **b**, The 383 *Hpa*II-*Hae*III fragment was used for S1 mapping analysis as described in Methods. The protected fragments are shown on the autoradiogram. Lane1, 200 U of S1 nuclease; lane2, 400 U of S1; lane (-S1), no S1 added. **c**, S1 mapping analysis of the 3′-end of exon 4 of *nad*4. The protected part of oligonucleotide O9 is shown in lane 1. The labelled O9 is shown in lane (-S1). Sequencing reactions were carried out using M13mp18 as size standard (lanes A,C,G,T). The products were resolved on a 6% polyacrylamide gel. The schemes under the autoradiograms give the coordinates (in parentheses) and refer to the DNA fragments shown in Figure 1a. The stars indicate the labelled extremities.

was performed in 10 μl of 10 mM Pipes pH 6.4, 0.4 M NaCl and 1 mM EDTA. After a denaturation step at 80°C for 3 min, the reaction mixture was incubated at 50°C for 1 h. The first strand synthesis was performed by addition of 90 μl of 50 mM Tris-HCl pH 8.3, 6 mM $MgCl_2$, 10 mM DTT, 0.4 mM of each dNTP and 20 U of AMV reverse transcriptase. The resulting cDNAs were amplified by PCR, using oligonucleotides O_1 and O_5, whose sequences were derived from the genomic DNA sequence modified in order to create *Xho* I sites at the 5'-extremities of both primers. The PCR amplification mix contained 16.6 mM $(NH_4)_2SO_4$, 67 mM Tris-HCl pH 8.8, 6.7 mM $MgCl_2$, 10 mM β-mercaptoethanol, 200 μM each dNTP, 0.1 mg/ml BSA, 30 mM KCl, 100 pmole of each primer and 4 U of *Taq* DNA polymerase (Pharmacia). Amplification was carried out by 30 cycles of the following steps : denaturation (94°C; 1 min); annealing (50°C; 2 min) and elongation (72°C; 3 min), followed by a final elongation step of 10 min at 72°C. Amplification products were phenol-extracted, ethanol-precipitated and purified on NACS columns (BRL). After digestion with *Xho* I, amplified cDNAs were cloned into the *Sal* I site of M13mp19 and sequenced using internal sequencing primers (O_3, O_4 and O_6, see Figure 1a and Table I).

RESULTS

Localization of exons 1, 2 and 3 of wheat mt *nad4*

A 820 bp *Bgl*II/*Hind*III probe containing part of the soybean gene coding for subunit IV of NADH dehydrogenase (8) was used to screen the complete *Sal*I library of wheat mt DNA as previously described (9). We have shown that only one clone, termed C3, corresponding to a 23 kb *Sal*I fragment of the wheat mt genome (17), hybridized with the soybean probe. In a previous

TABLE I. Synthetic oligonucleotides used for polymerase chain reaction, DNA and RNA sequencing and for Northern analysis.

		59	85
O_1	(D)	5'-AGCTCGAGTCTTTAGAAAGATTCCCGG-3'	
		197	172
O_2	(R)	5'-GGGAAGTCATGAGGCTAGCCCCCCC-3'	
		1914	1889
O_3	(R)	5'-GGATCCCAGTAAAGTATATAGGAAAA-3'	
		5873	5850
O_4	(R)	5'-TCCATAATATCTAACAAGTCGAGT-3'	
		6169	6143
O_5	(R)	5'-CCTCGAGCAAGAAAAGGTAGAAATATG-3'	
		2212	2240
O_6	(D)	5'-CCTTTCATTTATACTCTAAGCGCGATTGC-3'	
		6082	6105
O_7	(D)	5'-CTGGAAATTTAAAACCCGATTTCC-3'	
		7929	7901
O_8	(R)	5'-TTTCTTTTCTTTGCTGATTCCTCTCAATG-3'	
		7940	
O_9	(R)	5'-CCCGGAGAAGATTTCTTTTCTTTGCTGATT//	
		7894	
		//CCTCTCAATGAAATTTG-3'	

The orientation of the sequence as compared to the gene is indicated into parentheses, R: reverse, D: direct. Underlined numbers indicate the positions of the extremities of the oligonucleotides using the coordinates of wheat mt DNA sequence (Figure 1a). Dots indicate nucleotide mismatches introduced at the extremities of oligonucleotides O1 and O5 in order to create *Xho*I sites.

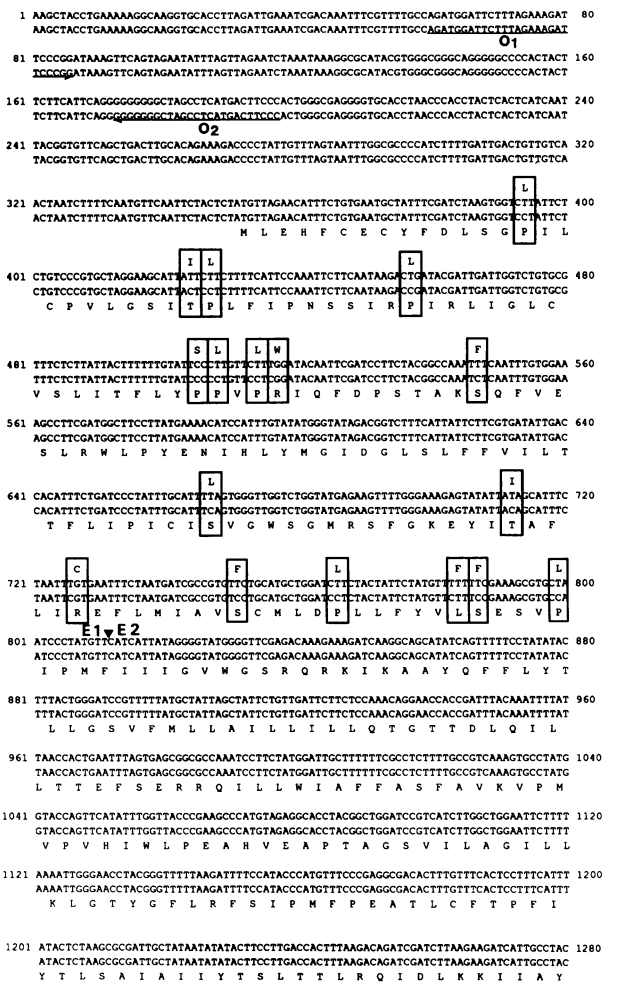

Figure 4. Comparison between wheat mitochondrial genomic DNA and cDNA sequences. The upper sequence corresponds to the cDNA sequence. The lower sequence is the corresponding genomic mt DNA sequence without intron sequences. The amino acid sequence deduced from the genomic sequence using the universal genetic code is shown below the corresponding sequence. Codons modified by RNA editing are boxed with the corresponding amino acid modification indicated above the cDNA sequence. Solid triangles indicate the position of the introns. The sequences of O_1 and O_5 (used for cDNA amplification), O_2 and O_8 (used for sequencing of direct, uncloned cDNA of respectively, the 5' non-coding region of E1 and E4) are underlined. The sequence of the *nad*4 cDNA has been deposited in the GenEMBL data bank under accession number X57163.

work, we have shown that the *nad*4 gene is interrupted in wheat mitochondria and have located two exons within the 1.9 kb and 2.9 kb *Sac*I fragments respectively (9). In order to determine the complete sequence of *nad*4, different *Sac*I fragments of C3, namely 1.9 kb, 0.4 kb, 2.9 kb and 2.1 kb (S19, S04,S29 and S21 in Figure 1a) have been sequenced. Comparison of the cDNA sequence (Figure 2b) with the genomic DNA sequence, shows the presence of a 3398 bp long intron (Figure 1a, I2). It was found that the upstream region of I2 is composed of an open reading frame of 225 codons preceeded by a number of a stop codons which indicated the presence of another intron.

The same strategy was used to cross the next intron boundaries. The cDNA sequence primed by oligonucleotide O_3 (Figure 1a, Table I) was determined (Figure 2a). The sequence of O_3 was deduced from the genomic DNA sequence about 50 bp downstream of the putative 5'-end of the open reading frame, estimated by the end of the aminoacid homology between the deduced wheat NAD4 protein sequence and its counterpart from *Aspergillus nidulans* (18). Comparison of the cDNA and genomic sequences shows that the 3'-end of the preceeding exon is located within the 2.1 kb *Sac*I fragment (Figure 1a, E1) and separated from the 5'-end of the dowstream exon by an intervening sequence of 1021 bp (Figure 1a, I1). Analysis of the sequence of this exon shows that an initiation codon can be found by homology with the other NAD4 proteins (Figure 6). At this point, the identified structure of the *nad*4 appeared to consist of three exons, E1, E2 and E3, interrupted by two introns I1 of 1021 bp and I2 of 3398 bp (Figure 1a).

Identification of the 5'-end of exon 1 and the 3'-end of exon 3

To determine whether the three identified exons correspond to the complete *nad*4 gene, we have localized the 5'-end of exon 1 and the 3'-end of exon 3 by primer extension and S1 mapping experiments, respectively. Primer extension experiments gave one predominant cDNA species of 196 bp (Figure 3a), indicating that the 5'-end of the transcript is situated 350 bp upstream of the putative AUG initiation codon. Other less intense signals could correspond either to less abundant transcripts or to premature termination of the reverse transcriptase reaction.

The 3'-end of E3 was localized by hybridization of the labelled coding strand of the 383 bp *Hpa*II-*Hae*III probe containing the putative 3'-end of exon 3 with total wheat mt RNA. An S1 nuclease protected fragment of 224 bp (position 6168 in Figure 3b) allowed the localization of the 3'-end of E3.

The structure of the gene is therefore as follows: exon 1 contains 811 bp (350 bp corresponding to the non-coding upstream region plus a 461 bp coding region). E2 contains 515 bp as deduced from the cDNA sequences used to identify the E1-E2 and E2-E3 junctions (Figure 2, a and b) and from the corresponding genomic DNA sequences (Figure 4). E3 contains 423 bp as deduced from the position of the E2-E3 junction (Figure 2b) and the location of its 3'-end (Figure 3b). Therefore, the three identified exons contain together 1749 bp, which is close to the size of the major *nad*4 transcript (1800 nt) observed by Northern hybridization (Figure 1b).

But the DNA sequence of E3 shows no termination codon (Figure 4). Moreover, the comparison of the protein sequence deduced from the three exons with other NAD4 proteins shows that the wheat mt protein would be 20 amino acids shorter than its human or *Aspergillus* counterparts. Since the transcript resulting from the three known exons is 1749 nt long, very close to the observed 1800 nt RNA (Figure 1b), it is possible that a termination codon could be created by an editing event in the 3'-end region of E3. However, only two arginine codons (*C*G-A→*U*GA, positions : 1422 and 1434) and two glutamine codons (*C*AG→*U*AG, position : 1335 and *C*AA→*U*AA, position : 1575) could be converted by RNA editing into a termination codon and the study of RNA editing of the *nad*4 transcript (Figure 4) shows that this is not the case. Therefore, we postulated the presence of a fourth exon.

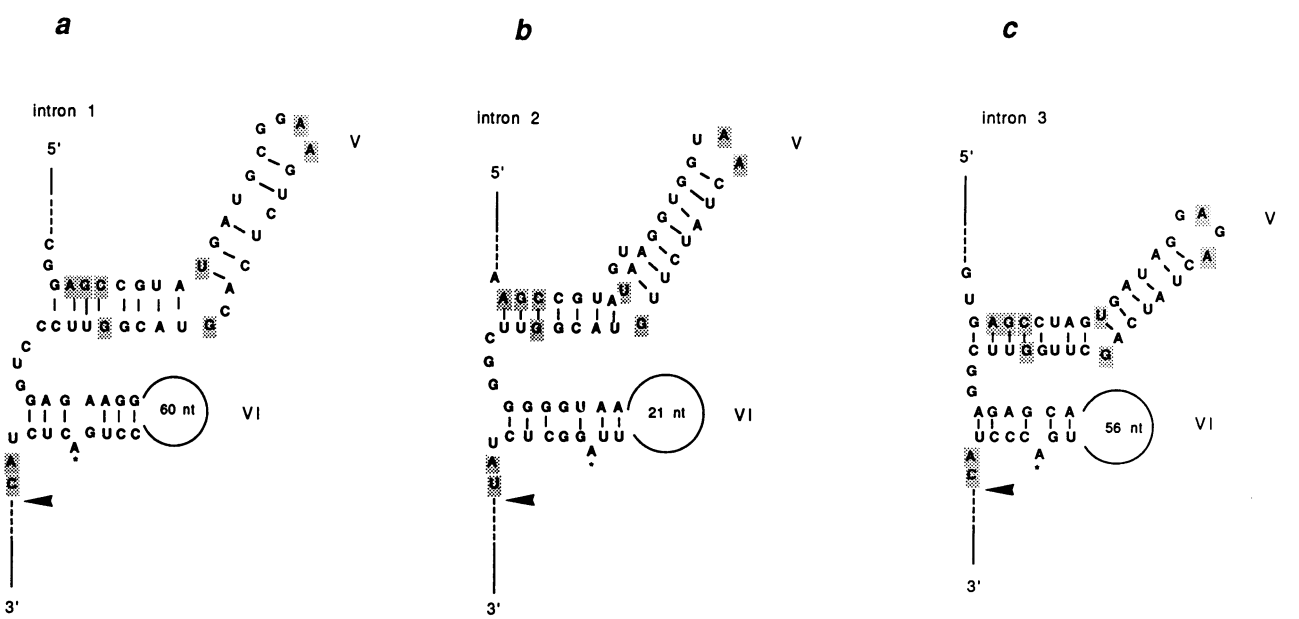

Figure 5. Consensus secondary structure of domains V and VI of the wheat mitochondrial *nad*4 introns. **a**, intron 1; **b**, intron 2 and **c**, intron 3. Shadowed nucleotides are quasi-universal (present in all group IIA introns, with at most one exception). Numbers represents the length of the variable single-stranded loops. The site of lariat formation is marked by an asterisk. Arrows point to the 3' intron-exon junctions of the three *nad*4 introns.

Searching for exon 4 (E4)

When the 3.4 kb *PstI* fragment (P34 in Figure 1a) was analyzed downstream of E3, it was possible to recognize a region which can be folded into conserved domains V and VI, a characteristic feature of organellar class II introns (Figure 5c). These secondary structures are present between 1543 and 1649 nucleotides downstream from the 3'-end of E3 previously determined by S1 mapping experiments (see Figure 3b). Therefore, if these assumptions are correct, the consensus terminal sequence of the putative 3'-end of intron 3 would be CCCUAC (Figure 5c) and the downstream nucleotides would constitute the 5'-end of exon 4 of *nad4*.

The nucleotide sequence of the fourth exon of *nad*4 was therefore determined (Figure 4). That this sequence is really exon 4 was confirmed by direct uncloned cDNA sequencing: O_8 (Figure 4) was annealed to total wheat mt RNA to prime a cDNA reaction using AMV reverse transcriptase (Figure 2c). This allows the crossing of exon 3-exon 4 boundary and the sequence obtained upstream was confirmed to be that of the 3' end of E3. This experiment confirms also the determination of the 3'-end of E3 as obtained by S1 mapping (see above). The 3'-end of E4 was determined by an S1 nuclease protection experiment using oligonucleotide O_9 (Figure 3c). After analysis of the peptide sequence deduced from the cDNA sequence, a termination codon was found 89 bp downstream of the E3-E4 junction in frame with the E3 deduced peptide sequence. The *nad*4 transcript extends 21 bp downstream of the termination codon. The *nad*4 mature transcript is 1859 nucleotides long. It contains 350 nt in the upstream non-coding region, 1485 nt from the initiation codon to termination codon, coding for a protein of 495 amino acids (Figure 4) and 21 nt in the downstream non-coding region.

Transcriptional analysis of the *nad4* gene

Two wheat mt RNA species are identified when internal probes corresponding to each of the four exons are used in Northern experiments (Figure 1b). All together, the four exons account for 1859 nt, the size of the most abundant transcript (Figure 1b). The same probes hybridize also to a larger transcript of about 2500 nt, which is present in very low amounts, about 100 times less than the 1859 nt transcript. We do not know the origin and the nucleotide sequence of the fragment which distinguishes the 1859 nt and the 2500 nt transcripts. Considering the sizes of I1, I2 and I3 (1.0 kb, 3.4 kb and 1.6 kb respectively), the 2500 nt transcript cannot results from the persistence of any of these introns. Since no significant open reading frame has been identified in I1, I2 or I3, the possibility of a maturase-type extension can also be ruled out.

Different probes corresponding to exons and introns of *nad*4 were used in Northern blot experiments. None of them detected products larger than 2500 nt, which could correspond to a primary non-spliced transcript of *nad*4 (not shown). When *in organello* transcription was performed using wheat mitochondria and (α-^{32}P)-UTP as precursor, the resulting labelled RNAs were shown to hybridize to the fragments containing the four exons of *nad*4 demonstrating that *nad*4 mRNA is produced in wheat mitochondria from the gene analysed here (in preparation). Since the primary transcript of *nad*4 is not detected in RNAs from green or etiolated wheat leaves, it must be rapidly processed.

Editing sites of the wheat *nad4* mRNA

It has already been shown that an editing site is located at the junction of E2 and E3 of the *nad*4 mRNA (9). Experiments have been undertaken to determine the total number of edited nucleotides in the 1859 nt long *nad*4 mRNA. cDNAs were synthesized and amplified using oligonucleotides O_1 and O_5 (Figure 4), corresponding to the extremities of E1 and E3, respectively. After DNAse treatment of the RNA preparations, cDNA synthesis was primed with O_5, complementary to the 3'-end of E3. No product was observed after PCR amplification when no reverse transcriptase was added during the cDNA reaction, demonstrating that no amplification was possible from any residual DNA. Moreover, the only DNA fragment obtained after amplification is 1.6 kb long (not shown), corresponding to the length of the three exons between O_1 and O_5, without any intron sequence, and cannot therefore be obtained from the DNA.

This 1.6 kb fragment was cloned into the *SalI* site of M13mp19 using the artificial *XhoI* sites present at the extremities of O_1 and O_5. Eight clones were analysed. All these clones are fully edited and contain 19 modifications when compared with the genomic DNA sequence (Figure 4). All these modifications are C to U conversions. Four editing sites have been identified on E4 by direct uncloned cDNA sequencing using O_8 as primer (Figure 4). E1 is edited at 17 positions, each of them modifying the nature of the encoded amino acid. This leads to a modification of 11% of the amino acid sequence genomically specified by E1. Surprisingly, E2 is not edited at all whereas E3 contains only 2 edited positions, one of them localized at the junction with E2 (Figure 4). E4 contains four edited nucleotides (two of them within the same codon) leading to the change of three amino acids. After edition, the deduced protein sequence corresponding to E4 is modified by 11%. No editing site was found in the upstream and downstream non-coding regions of the wheat mt *nad*4 transcript.

Similarity between different NAD4 proteins

The wheat mt NAD4 protein deduced from the cDNA nucleotide sequence is 495 amino acid long and has a molecular weight of 55812 Da with a calculated isoelectric point of 8.26. It shows 70%, 60% and 63% similarity with the corresponding proteins from *Aspergillus amstelodami* mitochondria (Nuo4$ASPAM in SwissProt protein data bank), mouse mitochondria (19) and tobacco chloroplast (20), respectively (Figure 6). Some of the domains (boxed in Figure 6) are better conserved, and may play a role in the function of the NADH subunit 4 or in the molecular interactions with other components of mitochondrial complex I.

About half of the 22 amino acid modifications resulting from RNA editing increases the similarity between wheat mitochondrial NAD4 protein and its counterparts from chloroplasts and non-plant mitochondria (Figure 6), leading to the incorporation either of the identical aminoacid (5 instances) or of an aminoacid which is functionnally equivalent (7 instances).

The partial DNA sequence of soybean mt *nad*4 (8) is homologous to that of E3 in wheat mt *nad*4 and is the only higher plant mitochondrial sequence available for comparison. If one compares the aminoacid sequence deduced from this partial soybean sequence with the corresponding part of the wheat protein, the similarity is 88%. The differences, at the DNA level, involve 17 nucleotides, twelve of them being C in soybean mt DNA and T in wheat mt DNA. They would all cause a change of the amino acid encoded by genomic DNA, with the exception of the tyrosine at position 390 (Figure 6). If these Cs were edited in the soybean transcript, the similarity between the two proteins would then increase to 96%. However, only four of these potential editing sites in soybean mt *nad*4 transcript would

enhance the conservation of the soybean protein as compared with non plant NAD4 proteins (Figure 6).

It is interesting to note that, if deduced from the genomic DNA sequence, the amino acid sequence of E1, which contains the majority of editing sites, shows the lowest similarity with the corresponding non-plant sequences, as compared with the other exons which are not (E2) or rarely (E3) edited.

DISCUSSION

Genomic structure of wheat mt *nad4*

This work describes the first complete plant mt *nad*4, its structural organization, its transcription and the determination of its editing sites. It has already been demonstrated (9) that this gene is present in only one copy in the wheat mitochondrial genome and it extends over a fragment of 8 kb.

The NAD4L gene, coding for another subunit of the mitochondrial NADH dehydrogenase complex (1), overlaps with the 5′ part of *nad*4 in the mitochondria of several species. This is true for mouse, human, bovine, *Drosophila melanogaster*, and *Xenopus laevis* mitochondria (19, 21−24). There is no evidence for the presence of an open reading frame homologous to *nad*4L in frame or out of frame in the wheat mt fragment containing *nad*4. Like watermelon and broad bean mt *nad*1 and *Oenothera* mt *nad*5 (3,4,7), the wheat mt *nad*4 gene is interrupted. It contains three introns which are respectively 1021, 3398 and 1649 bp long, intron 2 being the longest plant mt intron found up to now. It should be noted that no other *nad*4 from mammalian (19), fungal (18), protozoan mitochondria (25) or chloroplasts (20) contains introns. The search for consensus sequences at the 5′ and 3′ ends and the presence of consensus secondary structures in the 3′ part of these introns show that they belong to class IIA introns following the classification recently proposed (26).

Primary and secondary structures of *nad4* introns

When analyzed, the sequence of the three introns present in the *nad*4 gene can be partially folded to generate the highly conserved domains V and VI of group II introns (Figure 5 a, b and c) which is the group comprising plant mitochondrial introns (26). When the 3′ terminal part of the three introns are compared, one can find highly conserved sequences maintaining the elements of subgroup IIA of class II introns namely : All three introns end with YAY at their 3′extremities (Figure 5a, b and c), which is one of the features that distinguishes the subgroup IIA from the subgroup IIB. The bulging A on the 3′ side of helix VI is 7 nt upstream of the 3′ intron-exon junction for I1 and I3 and 9 nt for I2. The main diagnostic feature for class II introns (26) is that this bulging A, which participates in the lariat formation, is found at 7 or exceptionally 8 nucleotides upstream from the 3′ intron-exon junction. I2 possesses a G nucleotide at positions 7 and 8. However, it has already been demonstrated that correct lariat formation can be carried out if there is a G, instead of an A, at this position (27). No homology with maturase-like proteins (4) was found when internal intron sequences were translated and compared with the GenEMBL data bank. This fact and the presence of structural elements such as exon binding sites (EBS1 and EBS2), the bulging A residue and structural domain V, suggest that self-splicing mechanisms are involved in the processing of *nad*4 primary transcript.

Transcription of wheat mt *nad4*

In Northern experiments, every exon of *nad*4 hybridizes to a major 1800 nt RNA and also to a minor 2500 nt transcript. Determination of the 5′-end of the transcript by primer extension shows only one signal even after long exposure. From this point to the 3′-end of the transcript as determined by S1 mapping experiment (Figure 3c), the complete length of the transcript is 1859 nt which corresponds to the major signal observed by hybridization. Probes specific for the three introns of *nad*4 do not hybridize to the 2500 nt transcript indicating that it is not an unspliced precursor of the mature *nad*4 transcript. Moreover, when we tested the presence of *nad*4 transcripts in a wheat mitochondrial polysomal preparation, both the 1800 nt and the 2500 nt RNA species could be identified on Northern blots, in the same relative amounts as observed with total mt RNA

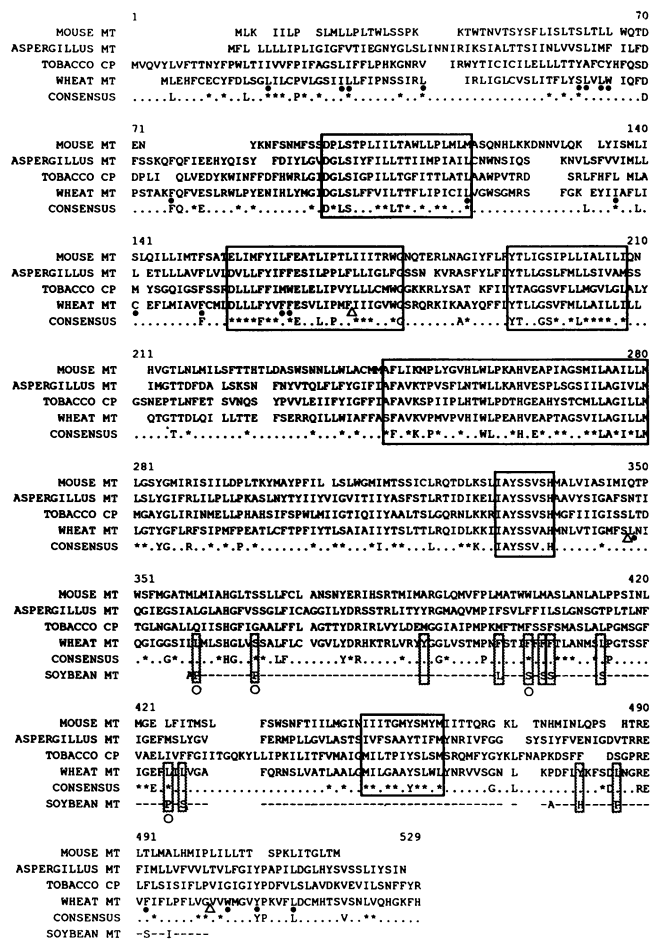

Figure 6. Amino acid similarities between wheat mitochondrial NAD4 protein and other NAD4 proteins. The wheat mt NAD4 protein deduced from cDNA is compared with the amino acid sequences (deduced from the gene sequences) of mitochondrial NAD4 proteins of mouse (19), *Aspergillus amstelodami* (see text), tobacco chloroplasts (20) and with the partial sequence of soybean mitochondria (8). Consensus sequence indicates: dots, the non-conserved positions; asterisks, functionnaly equivalent aminoacid; capital letters, complete identity. The regions of NAD4 proteins which are well conserved among species are indicated by solid boxes. Solid circles indicate the amino acids modified by editing in wheat mt *nad*4 transcripts. Empty triangles indicate the junction positions of wheat mt *nad*4 exons. A dash indicates identity to the amino acid deduced from wheat mt cDNA sequence. Dotted lines boxes indicate amino acids differences between soybean and wheat NAD4 proteins, which are due to the presence of a C in soybean genomic DNA and a T in wheat genomic DNA. Empty circles indicate putative editing sites of soybean mt *nad*4 transcripts, which would result in the incorporation of the same amino acid as specified by the wheat mt genomic sequence. The soybean amino acid sequence was taken from Figure 3 of reference 8.

preparation (not shown). This suggests that there is no precursor-product relationship between the two transcripts.

RNA editing of *nad4* transcript

The direct uncloned cDNA sequencing shows no ambiguity at the editing sites and this is consistent with the fact that all sequenced cDNA clones after PCR amplification were found fully edited. The genomic sequence of *nad4* differs from that of the mRNA (as deduced from the cDNA sequence) at 23 positions where RNA editing takes place (Figure 4). All these modifications are C to U conversions and result in a the change of the aminoacid encoded by the genomic DNA, with the possible exception of the editing site at position 1782 where a genomically encoded proline codon (*CC*G) is converted into a leucine codon (*UU*G) by two editing events. It is thus possible to consider the editing event at the first position of the codon as a silent one since CUG is also a leucine codon. Studies of this particular case of editing of the proline codons (CCN) into leucine codons (UUN) involving the modification of both adjacent Cs could well give some interesting insights about the mechanism of RNA editing.

No editing site was found in the non coding region of the *nad4* transcript. The editing sites are not evenly distributed along the mRNA and it is interesting to note that the great majority of editing sites are present in the first exon, while none could be found in the second exon. One can wonder whether such an uneven distribution of editing sites is related to the presence of functional domains in the protein or whether this is related to the hypothesis of exon shuffling, which proposes that, during evolution, each exon would act as an independent element in the constitution of different genes (28, 29) and therefore would be characterized by its own degree of RNA editing.

ACKNOWLEDGEMENTS

We thank Prof. J-H. Weil for critical reading of the manuscript and for his continual interest and support. We thank Dr Henri Wintz (Cornell University, Ithaca, USA) for providing the partial *nad4* soybean probe. We would also like to thank G.Bonnard, J.Canaday and D.Gonzalez (IBMP, Strasbourg) for their helpful comments during preparation of the manuscript. We thank also A.Koch for technical assistance. L.L. was supported by a fellowship first from CONICET (Argentina) and then from CNRS (France).

REFERENCES

1. Chomyn,A., Mariottini,P., Cleeter,M.W.J., Ragan,C.I., Matsuno-Yagi,A., Hatefi,Y., Doolittle,R.F. and Attardi,G. (1985) Nature **314**, 592–597.
2. Dujon,B. (1983) In Schweyen, R.J., Wolf, K. and Kaudewitz, F. (eds) Mitochondria 83, Walter de Gruyter, Berlin, New york, pp. 1–24.
3. Stern,D.B., Bang,A.G. and Thompson,W.F. (1986) Curr. Genet. **10**, 857–869.
4. Wahleithner,J.A., Macfarlane,J.L. and Wolstenholme,D.R. (1990) Proc. Natl. Acad. Sci. USA **87**, 548–552.
5. Xue,Y., Davies,D.R. and Thomas,C.M. (1990) Mol. Gen. Genet. **221**, 195–198.
6. Gualberto,J.M., Wintz,H., Weil,J.H. and Grienenberger,J.M. (1988) Mol. Gen. Genet. **215**, 118–127.
7. Wissinger,B., Hiesel,R., Schuster,W. and Brennicke,A. (1988) Mol. Gen. Genet. **212**, 56–65.
8. Wintz,H., Chen,H.C. and Pillay,D.T.N. (1989) Curr. Genet. **15**, 155–160.
9. Lamattina,L., Weil,J.H. and Grienenberger,J.M. (1989) FEBS Lett. **258**, 79–83.
10. Gualberto,J., Lamattina,L., Bonnard,G., Weil,J.H. and Grienenberger,J.M. (1989) Nature **341**, 660–662.
11. Covello,P.S. and Gray,M.W. (1989) Nature **341**, 662–666.
12. Hiesel,R., Wissinger,B., Schuster,W. and Brennicke,A. (1989) Science **246**, 1632–1634.
13. Stern,D.B. and Newton,K.J. (1984) Plant Mol. Biol. Reporter **2**, 8–15.
14. Maniatis,T., Fritsch,E. and Sambrook,J. (1982) Molecular cloning: A laboratory manual. Cold Spring Harbor Laboratory press, Cold Spring Harbor.
15. Sanger,F., Nicklen,S. and Coulson,A.R. (1977) Proc. Natl. Acad. Sci. USA **74**, 5463–5467.
16. Casanova,J.L., Pannetier,C., Jaulin,C. and Kourilsky,P. (1990) Nucleic Acids Res. **18**, 4028.
17. Quetier,F., Lejeune,B., Delorme,S., Falconet,D. and Jubier,M.F. (1985) In van Vloten-Doting, L., Groot, G.S.P. and Hall, T.C.(eds) Molecular form and function of the plant genomes. NATO ASI series A: life sciences 83, pp.413–420.
18. Brown,T.A., Davies,R.W., Ray,J.A., Waring,R.B. and Scazzocchio,C. (1983) EMBO J. **2**, 427–435.
19. Bibb,M.J., Von Etten,R.A., Wright,C.T., Walberg,M.W. and Clayton,D.A. (1981) Cell **26**, 167–180.
20. Shinozaki,K., Ohme,M., Tanaka,M., Wakasugi,T., Hayashida,N., Matsubayashi,T., Zaita,N., Chunwongse,J., Obokata,J., Yamaguchi-Shinozaki,K., Ohto,C., Torazawa,K., Meng,B.Y., Sugita,M., Deno,H., Kamogashira,T., Yamada,K., Kusuda,J., Takaiwa,F., Kato,A., Tohdoh,N., Shimida,H. and Sugiura,M. (1986) EMBO J. **5**, 2043–2049.
21. Anderson,S., Bankier,A.T., Barrell,B.G., De Bruijn,M.H.L., Coulson,A.R., Drouin,J., Eperon,I.C., Nierlich,D.P., Roe,B.A., Sanger,F., Schreir,P.H., Smith,A.J.H., Staden,R. and Young,I.G. (1981) Nature **290**, 457–465.
22. Anderson,S., De Bruijn,M.H.L., Coulson,A.R., Eperon,I.C., Sanger,F. and Young,I.G. (1982) J. Mol. Biol. **156**, 683–717.
23. Clary,D.O. and Wolstenholme,D.R. (1985) J. Mol. Evol. **22**, 252–271.
24. Roe,B.A., Ma,D.P., Wilson,R.K. and Wong,J.F.H. (1985) J. Biol. Chem. **260**, 9759–9774.
25. Benne, R. (1990) Trends Genet. **6**, 177–181.
26. Michel,F., Umesono,K. and Oseki,H. (1989) Gene **82**, 5–30.
27. Schmelzer,C. and Muller,M.W. (1987) Cell **51**, 753–762.
28. Blake,C. (1983) Nature **306**, 535–537.
29. Dorit,L.R., Schoenbach,L. and Gilbert,W. (1990) Science **250**, 1377–1382.