Vol. 119, No. 2, 1984 BIOCHEMICAL AND BIOPHYSICAL RESEARCH COMMUNICATIONS
March 15, 1984 Pages 786-794

IDENTIFICATION OF THE PROBABLE CODING REGION FOR EXON 2 OF CYTOCHROME OXIDASE POLYPEPTIDE I FROM ASPERGILLUS NIDULANS MITOCHONDRION

David G. George, Lai-Su L. Yeh, Lois T. Hunt, and Winona C. Barker

National Biomedical Research Foundation
Georgetown University Medical Center
3900 Reservoir Road, N.W.
Washington, D.C. 20007

Received February 7, 1984

SUMMARY: Hypothetical protein URFe of Aspergillus nidulans mitochondrion is homologous with the amino end of cytochrome oxidase (EC 1.9.3.1) polypeptide I. Unidentified reading frame URFe does not contain a suitable initiation codon and codes for a protein with a length of only 91 residues, corresponding to about 20% of cytochrome oxidase polypeptide I. It is proposed that this region codes for the second exon of the cox1 gene of Aspergillus mitochondrion. Possible candidates for the 2- to 3-residue amino-terminal exon 1 are discussed.

Netzker et al. (1) have reported the nucleotide sequence of about 44% of the Aspergillus nidulans mitochondrial genome. This region contains the genes for ATPase (EC 3.6.1.3) subunit 6, cytochrome oxidase (EC 1.9.3.1) polypeptide III, and four tRNAs, as well as seven unidentified reading frames (URFs). Comparison of the peptides translated from these coding regions with other sequences in our Protein Sequence Database (2,3) indicated that the sequence coded by URFe was similar to the amino end of cytochrome oxidase polypeptide I. URFe occurs within a 1-kilobase region between the His tRNA gene and the 3' end of the sequence reported by Netzker et al. (1). No known genes occur within this region.

Program ALIGN (3) was used to evaluate the statistical significance of the relationship between hypothetical protein URFe and the amino end of baker's yeast (Saccharomyces cerevisiae) cytochrome oxidase polypeptide I. The comparison of these two peptides obtained an ALIGN score greater than 21 SD; the probability of obtaining an ALIGN score of 10 SD units or greater by chance alone is less than 10^{-23}. Figure 1 shows an alignment of the Aspergillus sequence with the amino ends of cytochrome oxidase polypeptides I from yeast,

0006-291X/84 $1.50

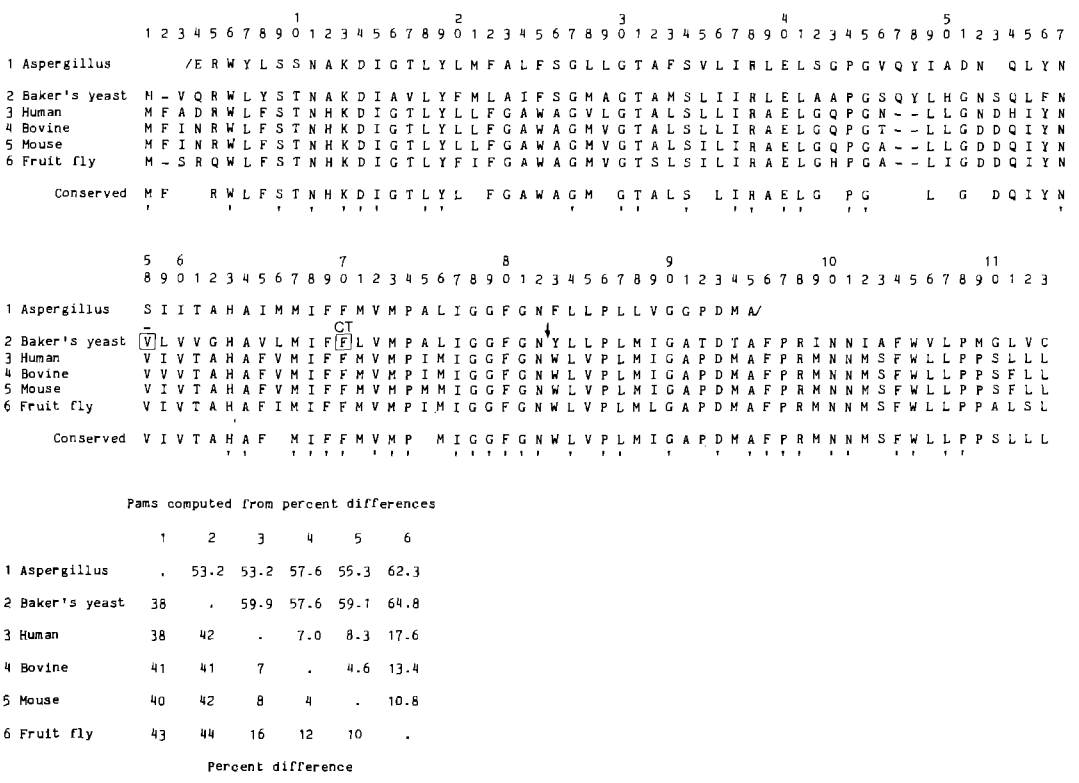

Figure 1. Alignment of hypothetical protein URFe from Aspergillus nidulans mitochondrion (1) with the amino end of cytochrome oxidase polypeptide I from baker's yeast (4), human (8), bovine (9), mouse (10), and fruit fly (11) mitochondria. The codons corresponding to the residues in the yeast sequence that are enclosed by boxes are split by the first two introns of the yeast oxi3 gene; the amino acids directly above the boxes are those assigned by Bonitz et al. (4). The position of the third intron in the yeast sequence is indicated by the arrow.

human, bovine, mouse, and fruit fly mitochondria. Strong similarity can be seen between the Aspergillus sequence and all of the known sequences of cytochrome oxidase polypeptide I.

The oxi3 (cytochrome oxidase polypeptide I) gene of yeast contains seven introns (4), three of which occur within the region homologous with URFe. Figure 2 shows an alignment of the 5' end of the oxi3 gene and URFe. The introns have been removed from the yeast sequence; their positions are indicated by diagonal lines in the figure. The nucleotide sequences have been highly conserved; the regions beginning at about the third codon of the yeast protein and ending 90 codons later, near a pair of TAA termination codons in the Aspergillus sequence, are only 27% different. The amino end of the

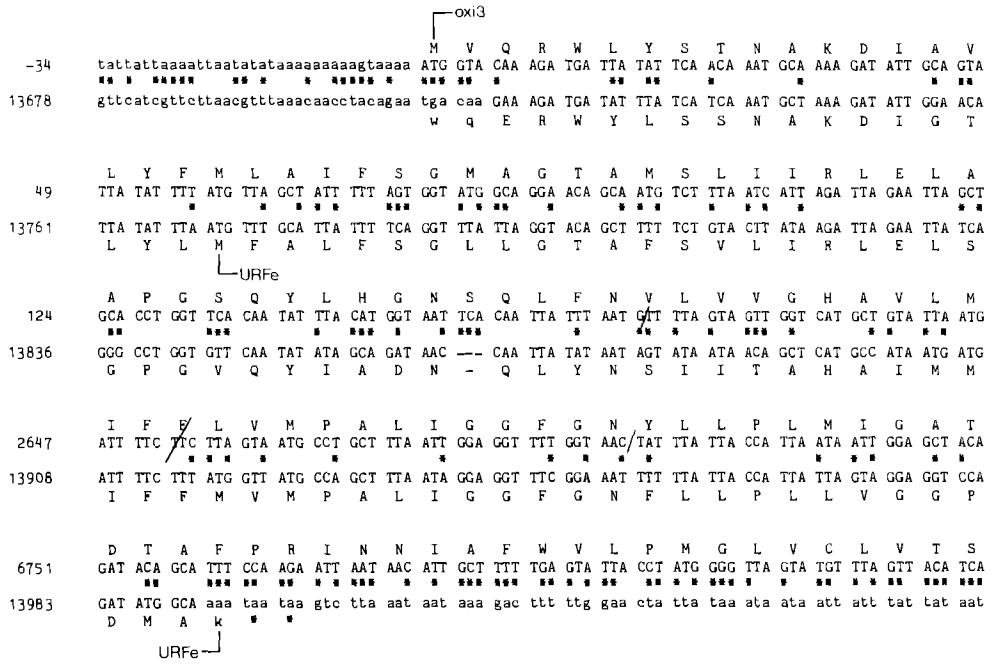

Figure 2. Alignment of the oxi3 gene coding region of baker's yeast mitochondrion (4) and the homologous region from Aspergillus nidulans mitochondrion (1). The yeast sequence is shown on top. The positions of the three introns in the yeast sequence (residues 170-2617, 2654-5168, and 5207-6720) are denoted by diagonal lines. The coding region of the yeast oxi3 gene and the proposed coding region for exon 2 of the Aspergillus cox1 gene are indicated by uppercase letters; protein sequences are indicated above and below the coding regions. The original numbering systems of the authors are shown.

Aspergillus sequence is difficult to identify; there is no formylmethionine codon in the region corresponding to the amino end of the yeast sequence. The first formylmethionine codon occurs at the 20th codon; this codon corresponds to the 5' end of URFe as defined by Netzker et al. (1). The open reading frame actually begins about 40 codons before the first codon shown in Figure 2, and there is no formylmethionine codon in this entire region. Furthermore, the two TAA codons in the Aspergillus sequence truncate the protein at 91 residues, far short of the 512-514 residues of cytochrome oxidase polypeptide I.

There is little doubt that URFe is homologous with the 5' end of the yeast oxi3 gene. Without direct experimental confirmation, however, it cannot be demonstrated that URFe is a portion of the functional cox1 (cytochrome oxidase polypeptide I) gene of Aspergillus. Let us assume that URFe is a fragment of the cox1 gene; then the simplest interpretation is that URFe corresponds to

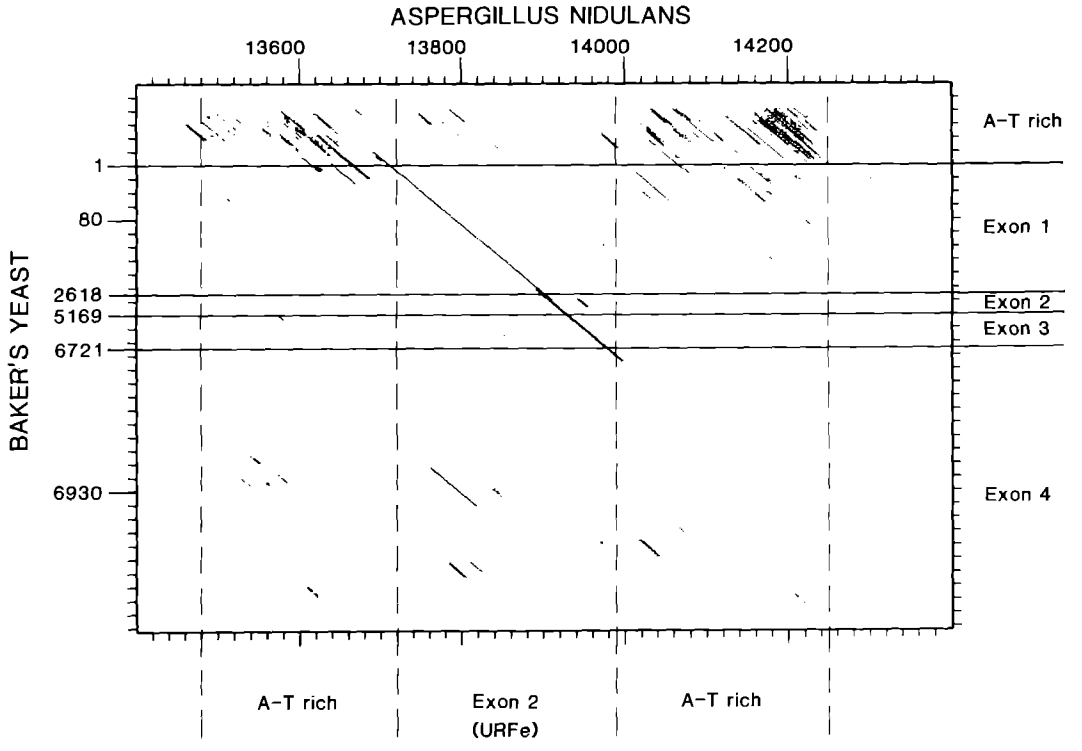

Figure 3. DOTMATRIX (5) comparison of the nucleotide sequence of the region surrounding URFe (1) with the 5' end of the coding region of the yeast oxi3 gene (4). Segments of length 75 were compared using a unitary nucleic acid scoring matrix and a minimum scoring value of 38; this corresponds to about 50% identity. The original numbering systems of the authors are shown.

the second exon of this gene. Although introns are not present in animal mitochondrial genes, they are observed in the mitochondria of fungi and plants. The genes for apocytochrome b and 21S rRNA of Aspergillus both contain introns.

A DOTMATRIX comparison (5) of the nucleotide sequence surrounding URFe and the 5' end of the coding region of the yeast oxi3 gene is shown in Figure 3. Sequence similarity is indicated by diagonal lines in the plot. The longest diagonal corresponds to the regions of observed sequence similarity between URFe and the yeast oxi3 gene. This diagonal line ends in the vicinity of the two TAA termination codons in the Aspergillus sequence, indicating that the similarity does not extend beyond these positions. Thus, the third exon of cox1 is probably not contained in the region sequenced by Netzker et al. (1).

As shown in Figure 3, the region immediately preceding the oxi3 gene contains segments that are similar to segments in the regions 5' and 3' to URFe. The region preceding the oxi3 gene is AT-rich (95.3% A + T). Although both the yeast and the Aspergillus mitochondrial genomes have a high A + T content (the region reported by Netzker et al. [1] contains 74% A + T), the intervening regions show enhanced A + T content relative to the coding regions (6). The regions immediately preceding and following URFe contain 81.3% and 84.0% A + T whereas URFe contains 70.6% A + T. Thus, the assignment of these two regions as introns is in accord with observed nucleotide frequencies. The relatively high A + T contents account for the similarity observed between these regions and the region preceding the yeast gene.

Assuming URFe corresponds to exon 2 of the cox1 gene, it is of interest to define precisely the boundaries of this exon. Based on the nucleotide sequences of introns found in cytoplasmic genes, two consensus sequences have been proposed to define the 5' and 3' intron junctions (7). Figure 4 shows the boundaries of the seven introns of the yeast mitochondrial oxi3 gene. It can readily be seen from this figure that only the junctions of introns 1 and 2 show any similarity to the consensus sequences defined by the cytoplasmic genes. Thus, the consensus sequences cannot be used to locate the mitochondrial intron junctions.

It should be noted that the boundaries shown in Figure 4 for introns 1, 2, and 5 differ from those assigned by Bonitz et al. (4). The introns given by these authors are indicated by lowercase letters in the figure. Comparison of the yeast sequence with the corresponding protein sequences of human, bovine, mouse, and fruit fly mitochondria indicated that the junctions should be modified. Intron 1 occurs at position 58 in Figure 1. The assignment of Bonitz et al. (4) results in a gap in the protein sequence at this position. Because all of the corresponding sequences have Val at position 58, it is more likely that the intron occurs between nucleotides 169 and 2618, yielding the codon GTT for Val at this position in the yeast sequence. It is of interest that these newly defined boundaries, TG|GTGCG and CTATTTCAT|TT, are in reasonably

Intron Junctions of the Yeast oxi3 Gene

*Intron 1	170-2617	F N TTT AAT g	gtgcgcctctc...	...gctatttcat	V L V tt TTA GTA
*Intron 2	2654-5168	I F ATT TTC T	GTgcgccgttt...	...gctactctaC	F L V TC TTA GTA
Intron 3	5207-6720	F G N TTT GGT AAC	caaaaaagata...	...taaaatgaac	Y L L TAT TTA TTA
Intron 4	7198-8207	F F G TTC TTT GGT	caaacagtggc...	...atataacaag	H P E CAC CCT GAA
*Intron 5	8620-9506	H F CAT TTT c	gagcggtctga...	...ctatcgggaT	H Y V AC TAT GTA
Intron 6	9809-9829	L V N TTA GTT AAT	ggattaaataa	taaagttaat	N K S AAT AAA TCA
Intron 7	9860-9904	K A P AAA GCA CCC	gattttgtaga...	...tacagttaaa	S S S TCT TCA TCT
Consensus Intron Junctions			AG gtrag	ynyyyncag	

Proposed Intron Junctions of the Aspergillus cox1 Gene

Intron 1	13547-13717	M Y S ATG TAT AG	catacgtaaag...	...cagaatgaca	E R A GAA AGA
Intron 1	13548-13718	M Y S ATG TAT AGC	atacgtaaagt...	...agaatgacaa	E R W GAA AGA TGA
Intron 2	13992-	D M A GAT ATG GCA	aaataataagt...		

Figure 4. The intron junctions of the yeast oxi3 gene (4) and the proposed intron junctions of the Aspergillus cox1 gene (1). The junctions of the yeast introns marked by asterisks differ from those assigned by Bonitz et al. (4); the introns specified by these authors are shown in lowercase. Two alternative junctions are shown for intron 1 of the cox1 gene. The original numbering systems of the authors are shown. The consensus intron junction sequences are those defined by Sharp (7). R = A or G; Y = C or T; N = A, C, G, or T.

good agreement with the consensus intron sequences. Intron 2 occurs at position 70 in the alignment shown in Figure 1. The assignment indicated in Figure 4 produces Phe at this position rather than Cys-Thr as proposed by Bonitz et al. (4). All of the other corresponding sequences, including the one from Aspergillus, have Phe at this position. Moreover, the region surrounding this residue is highly conserved; the three preceding residues are absolutely conserved and, with the exception of the Leu at position 71 of the yeast sequence, the following four residues are absolutely conserved. This assignment makes the boundaries of this intron also more closely resemble the consensus intron sequences. The boundaries for intron 5 indicated in Figure 4

produce a CAC codon for His rather than a TAC codon for Tyr as defined by Bonitz et al. (4). All other sequences have His at this position, which occurs within a stretch of 10 absolutely conserved residues (data not shown). The junctions of introns 3 and 4 would more closely resemble the consensus intron sequences if they were shifted. However, both of these introns occur in conserved regions of the sequence and altering the boundaries would result in protein dissimilarities at these positions. Thus, the consensus intron sequences alone do not provide a good indication of the mitochondrial intron boundaries.

The proposed coding region for exon 2 of the Aspergillus cox1 gene is indicated in uppercase in Figure 2. The last four amino acids coded by this region, Pro-Asp-Met-Ala, are conserved in the sequences from human, bovine, mouse, and fruit fly. The yeast sequence differs at two of these positions. The next codon of the Aspergillus sequence codes for Lys, followed by the two termination codons; all other sequences have Phe-Pro-Arg at the corresponding positions. Furthermore, as indicated in Figure 2, the nucleotide sequences of the Aspergillus and yeast genes become highly dissimilar beginning at the Aspergillus AAA codon for Lys. Thus, it is likely that exon 2 ends at the GCA codon at nucleotides 13989-13991.

As indicated in Figure 1, exon 1 of the cox1 gene is expected to code for a peptide of two or three amino acids. There are six ATG codons in the region between the 3' end of the coding region for His tRNA and the proposed coding region for exon 2 of the cox1 gene; they are shown in Figure 5. Because of the small size of exon 1, none of these six possible candidates can be completely ruled out. Two of the ATG codons are immediately followed by TAA termination codons, and one is followed by a codon for Tyr, which is immediately followed by a TAG termination codon. The ATG codon at 13712 is only eight residues before URFe and thus would require a translational frameshift or an unreasonably small intron in order to serve as the amino end of the cox1 gene. The best candidate for exon 1 is the segment beginning at nucleotide 13539; this segment codes for the peptide Met-Tyr-Ser, which aligns very well with

```
                 M   *
13441-13446      ATG TAA

                 M   Y   S   I   R   K   V   C   Y   *
13539-13568      ATG TAT AGC ATA CGT AAA GTA TGT TAT TAA

                 M   L   L   M   *
13559-13573      ATG TTA TTA ATG TAA

                 M   *
13568-13573      ATG TAA

                 M   Y   *
13618-13626      ATG TAT TAG

                 M   T   r   k   m   i   f   i   i   k   c   *
13712-13747      ATG ACA Aga aag atg ata ttt atc atc aaa tgc taa
                         E   R   W   Y   L   S   S   N   A   K
```

Figure 5. Possible coding regions for exon 1 of the Aspergillus cox1 gene. All ATG codons within the region from the end of the His tRNA coding region (nucleotide 13428) to the proposed 5' end of exon 2 (nucleotide 13719) are shown. The segment from 13712-13747 contains the 5' end of exon 2, which is shown in lowercase; its translation is shown below the nucleotide sequence.

the amino end of the proteins shown in Figure 1. Tyr-2 aligns with Phe-2 of the mammalian mitochondrial sequences; a mutation from Tyr to Phe requires only one base change, and Tyr has empirically been found to replace Phe quite frequently (3). Ser-3 would align with Ser-3 of the fruit fly sequence. The proposed sequence of exon 1 and the intron junctions are shown in Figure 4. There are two alternate pairs of intron junctions that produce the same sequence at the amino end of the cox1 protein. Neither of these alternate junctions is similar to the consensus intron junction (7) and both are equally good candidates.

NOTE: In the final preparation of this manuscript it came to our attention that F. Michel and B. Dujon (EMBO J. 2, 33-38, 1983) have also found it necessary to redefine the intron junctions of introns 1, 2, and 5 of the yeast oxi3 gene. Their redefinition was based upon compliance with secondary structure models for group II introns. They reported new junctions for the 5' and 3' ends of introns 1 and 2 and for the 3' end of intron 5. These junctions are identical with the junctions that we have proposed.

ACKNOWLEDGMENTS: This research was supported by NIH grant GM-08710 from the National Institute of General Medical Sciences and by NASA Contract NASW 3317.

REFERENCES

1. Netzker, R., Kochel, H.G., Basak, N., and Kuntzel, H. (1982) Nucl. Acids Res. 10, 4783-4794.
2. Orcutt, B.C., George, D.G., and Dayhoff, M.O. (1983) Annu. Rev. Biophys. Bioeng. 12, 419-441.
3. Dayhoff, M.O., Barker, W.C., and Hunt, L.T. (1983) Methods Enzymol. 91, 524-545.

4. Bonitz, S.G., Coruzzi, G., Thalenfeld, B.E., Tzagoloff, A., and Macino, G. (1980) J. Biol. Chem. 255, 11927-11941.
5. George, D.G., Yeh, L.-S.L., and Barker, W.C. (1983) Biochem. Biophys. Res. Commun. 115, 1061-1068.
6. Macino, G., and Morelli, G. (1983) J. Biol. Chem. 258, 13230-13235.
7. Sharp, P.A. (1981) Cell 23, 643-646.
8. 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., Schreier, P.H., Smith, A.J.H., Staden, R., and Young, I.G. (1981) Nature 290, 457-465.
9. 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.
10. Bibb, M.J., Van Etten, R.A., Wright, C.T., Walberg, M.W., and Clayton, D.A. (1981) Cell 26, 167-180.
11. de Bruijn, M.H.L. (1983) Nature 304, 234-241.