The EMBO Journal Vol.2 No.3 pp.427–435, 1983

The mitochondrial genome of *Aspergillus nidulans* contains reading frames homologous to the human URFs 1 and 4

Terence A. Brown*, R. Wayne Davies, John A. Ray, Richard B. Waring and Claudio Scazzocchio[1]

Department of Biochemistry, University of Manchester Institute of Science and Technology, Manchester M60 1QD, and [1]Department of Biology, Essex University, Wivenhoe Park, Colchester C04 3SQ, UK

Communicated by L.A. Grivell
Received on 7 January 1983

A 2830-bp segment of the mitochondrial genome of the fungus *Aspergillus nidulans* was sequenced and shown to contain two unidentified reading frames (URFs). These reading frames are 352 and 488 codons in length, and would specify unmodified proteins of mol. wts. 39 000 and 54 000, respectively. The derived amino acid sequences indicate that these genes are equivalent to the human mitochondrial URFs 1 and 4, with 39% amino acid homology for URF1 and 26% for URF4. Both URFs were shown by secondary structure predictions to code for predominantly β-sheeted proteins with strong structural conservation between the fungal and human homologues. Counterparts of mammalian URFs have not previously been identified in non-mammalian genomes, and the discovery that *A. nidulans* possesses reading frames so closely homologous with URF1 and URF4 shows that these genes are of general functional importance in the mitochondria of diverse species.

Key words: Aspergillus nidulans/DNA sequencing/mitochondrial DNA/unidentified reading frames

Introduction

Before the advent of DNA sequencing the study of the mitochondrial genome was in general limited to genetic analysis of lower organisms, in particular *Saccharomyces cerevisiae*. Six, possibly seven, protein-coding genes had been located on the yeast map, with the genome saturated with genetic loci to such an extent that it was thought unlikely that any additional structural genes were present (Tzagoloff *et al.*, 1979). However, the complete human mitochondrial DNA sequence (Anderson *et al.*, 1981) revealed a total of 13 possible protein-coding genes. Five of these were counterparts of known yeast genes, leaving eight unidentified reading frames (URFs), each transcribed into RNA products (Montoya *et al.*, 1981; Ojala *et al.*, 1981), and each capable of coding for a substantial polypeptide. The sequences of mouse (Bibb *et al.*, 1981) and bovine (Anderson *et al.*, 1982) mitochondrial genomes also proved to possess reading frames homologous to the human URFs, indicating that these genes are standard features of the mammalian mitochondrial genome.

By 1981 considerable parts of the DNA of the *S. cerevisiae* mitochondrial genome had been sequenced, and it was clear that no counterparts of the human URFs were present in the regions completed at that stage. Although limited regions of the yeast genome still remain uncharacterised, the failure of further studies to locate reading frames equivalent to mammalian URFs raises the possibility that these genes are not present in yeast mitochondrial DNA. This absence could be explained in one of two ways. The first possibility is that homologues of the human URFs are nuclearly-encoded in yeast, with translation products transported into mitochondria after synthesis. The alternative to nuclear URFs is that these genes are not universal and are only found in mammals.

Apart from the three mammalian genomes, the most completely sequenced mitochondrial DNA is that of the ascomycete *Aspergillus nidulans*. During the last 3 years two laboratories, ours and that of H. Küntzel, have sequenced ~27 kb of this 32.5-kb circular molecule, and an accurate map of the genome can now be drawn (Figure 1). Of the eight URFs found so far, two display clear sequence homology with human URFs (Davies *et al.*, 1982). These reading frames, located on the *A. nidulans* map downstream of the apocytochrome b gene (*cob*A), are unmistakably the counterparts of URF1 and URF4, and are therefore the first examples of homologues of mammalian mitochondrial URF genes.

The discovery of URF1 and URF4 in the mitochondrial genome of *A. nidulans* shows that these genes are important in the mitochondria of widely divergent species, and facilitates the characterisation of the URF gene products, using the nuclear and the mitochondrial genetics systems that have been developed with this organism (Waring and Scazzocchio, 1982). Complete sequence characterisation of the genes is an essential preliminary to such an analysis. Here we present the DNA sequence of the region of the *A. nidulans* genome that contains URF1 and URF4, and provide a comparison of the fungal and human URF gene products at the levels of primary amino acid sequence and polypeptide secondary structure.

Whilst this paper was in preparation two reports emphasised further the general importance of URF1. Firstly, hybridisation experiments with *A. nidulans* probes have located URF1 in the mitochondrial genomes of tobacco and maize (Waring *et al.*, in preparation). Secondly, the start of a gene homologous to URF1 is present in a recently published sequence of *Drosophila melanogaster* mitochondrial DNA (Clary *et al.*, 1982).

Results

Codon usage in URF1 and URF4 is typical of the exon regions of A. nidulans mitochondrial genes

The strategy used to sequence URF1 and URF4 is described in the legend to Figure 2, and the complete sequence of this region is presented in Figure 3. The complete sequence confirms previous identifications of URF1 and URF4 based on partial sequence data (Davies *et al.*, 1982; Netzker *et al.*, 1982).

The URF1 and URF4 reading frames are of 352 and 488 codons, and would be translated into unmodified proteins of mol. wts. 39 000 and 54 000, respectively. All five other possible reading frames are multiply closed in each gene (see legend to Figure 3).

The pattern of codon usage in URF1 and URF4 is similar to that observed for exon regions of other mitochondrial genes in the *A. nidulans* genome (Waring *et al.*, 1981; Net-

*To whom reprint requests should be sent.

© IRL Press Limited, Oxford, England.

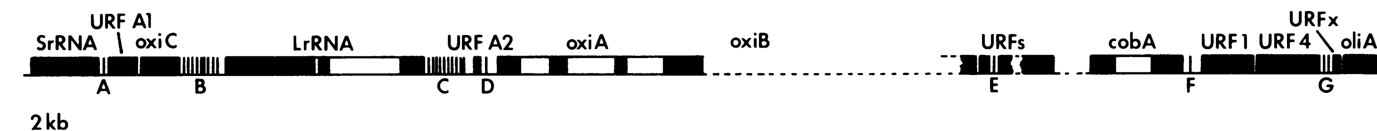

Fig. 1. Genetic organisation of the *A. nidulans* mitochondrial genome. The circular genome is drawn as a linear map with the positions of five protein-coding genes, eight URFs, 23 tRNAs and two rRNA genes marked. The URFs exclude reading frames present in the introns of *oxi*A and *cob*A which, although unidentified, are believed to code for maturases (Waring *et al.*, 1982 and in preparation). Complete nucleotide sequence data exists for all but the dotted portion of the genome (Köchel and Küntzel, 1981; Köchel *et al.*, 1981; Waring *et al.*, 1981, 1982; Grisi *et al.*, 1982; Netzker *et al.*, 1982; Waring *et al.*, in preparation; this paper plus unpublished data). In previous papers the URF nomenclature has been confused: we propose to retain the names 'URF1' and 'URF4' for those reading frames equivalent to mammalian URFs, and to specify reading frames unique to *A. nidulans* by URF A1, URF A2, etc. An ambiguity exists with 'URFx' which possesses limited homology to the mammalian reading frame URF A6L (Grisi *et al.*, 1982; Netzker *et al.*, 1982). Abbreviations: SrRNA, 16S rRNA; LrRNA, 23S rRNA; *oxi*A, B and C, cytochrome oxidase subunits 1, 2 and 3; *cob*A, apocytochrome b; *oli*A, ATPase subunit 6. The tRNA genes are indicated by letters: A, $tRNA^{tyr}$; B, upstream cluster (tRNAs for lys, $gly^{1,2}$, asp, $ser^{1,2}$, trp, ile and pro); C, downstream cluster (thr, glu, val, $met^{1,2,3}$, $leu^{1,2}$, ala, phe, gln); D, $tRNA^{his}$; E, $tRNA^{asn}$; F, $tRNA^{cys}$; G, $tRNA^{arg}$ and $tRNA^{asn}$. The LrRNA, *oxi*A and *cob*A introns are shown as open areas within these genes.

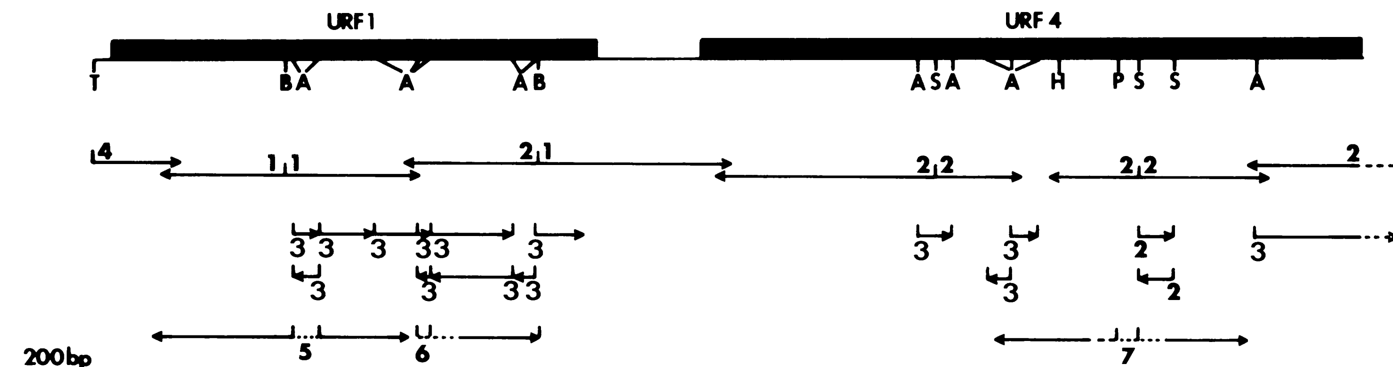

Fig. 2. Sequencing strategy for the region containing URF1 and URF4. URF1 was located from 5′-terminal sequences of three M13mp6 clones of total mitochondrial DNA double-digested with *Bgl*II and *Bcl*I. The positions of these sequences, which showed clear homology with human URF1 (Anderson *et al.*, 1981), are shown by the arrows labelled 1. at the tails. Additional sequencing of subclones derived from pBR322-*Bgl*II fragment 3 clone provided more of the URF1 gene and also revealed that this gene is immediately followed by URF4 on the genome map. Arrows representing these subclones are labelled as follows: 2. *Sau*3A subclones; 3. *Alu*I subclones; 4. *Taq*I subclones. The sequence was completed by preparing three restriction fragments for use as internal primers onto the relevant *Bgl*II-*Bcl*I clone. Sequences obtained with these primers are shown by arrows labelled: 5. 42-bp *Alu*I fragment; 6. 27-bp *Alu*I fragment; 7. 48-bp *Pvu*II-*Sau*3A fragment. All sequences were read in the direction of the arrows. Restriction sites: A, *Alu*I; B, *Bcl*I; H, *Hin*dIII; P, *Pvu*II; S, *Sau*3A; T, *Taq*I.

zker *et al.*, 1982) with a strong bias towards codons ending in A or U, rather than G or C. This pattern differs from the codon usage employed by the reading frames present in the introns of the genes for apocytochrome b (Waring *et al.*, 1982) and cytochrome oxidase subunit 1 (Waring *et al.*, in preparation) and indicates that URF1 and URF4 are typical protein-coding genes.

The derived amino acid sequences reveal homology between fungal and mammalian URFs

Designation of the two reading frames as URF1 and URF4 is justified by comparison of the predicted amino acid sequence of the *A. nidulans* gene products with the human counterparts. Alignments between the *A. nidulans* and human gene products, using the minimum number of sequence discontinuities that are consistent with obvious regional conservation, result in amino acid homologies of 39% for URF1 and 26% for URF4. Those residues that are conserved between the two species are marked by asterisks in Figure 3. The values for the URFs are comparable with the degree of amino acid conservation between known protein-coding genes in the fungal and human genomes, examples being 51% for apocytochrome b (Waring *et al.*, 1982), 49% for cytochrome oxidase subunit 3 (Netzker *et al.*, 1982) and 32% for ATPase subunit 6 (Grisi *et al.*, 1982). A higher level of amino acid conservation is detectable when the *A. nidulans* URFs are compared with the equivalent reading frames from the bovine (Anderson *et al.*, 1982) and mouse (Bibb *et al.*, 1981) genomes (Figure 4). Figures of 46% for URF1 and 32% for URF4 are obtained when a conserved residue is defined as one present in *A. nidulans* and at least one of the mammalian species.

Although the alignment used to compare the URF gene products from different species requires very few discontinuities, there are regions where gaps occur in one or other sequence. In particular, the region of URF1 bounded by the conserved peptides FLG and KT, and underlined in Figure 3, includes 42 codons in the *A. nidulans* sequence but only 17 in the mammalian genes. There is no recognisable homology between *A. nidulans* and the human genes in this region which is also variable amongst the mammalian URFs, with only six out of 17 residues conserved between the human, mouse and bovine genes (Figure 4), giving a homology of 35% for this segment compared with 78% when the entire mammalian URF1 sequences are considered. The nucleotide sequence of the fungal gene in this region has an AT content of 86%, significantly greater than elsewhere in either URF1 and URF4. Indeed, it is quite unusual to find a sequence of 126 nucleotides with such a high AT content within the coding regions of *A. nidulans* mitochondrial DNA, a high

```
TCGAATAAAGCATTAATATAAAGTTATTTGTATTTTTACAAAGATATTATAAAAAAATGT
                                             START URF 1  M
                                        100
TATATTTAATTTCAATAATTGAAGGTCTTTTAGTAATTGTTCCTGCTTTATTATCAGTTG
L  Y  L  I  S  I  I  E  G  L  L  V  I  V  P  A  L  L  S  V
                           *  *     *  *  *     *

CTTTTGTAACAGTAGCAGAAAGAAAAACTATGGCAAGTATGCAAAGAAGATTAGGTCCTA
A  F  V  T  V  A  E  R  K  T  M  A  S  M  Q  R  R  L  G  P
*  *              *  *  *              *  *     *     *  *
                 200
ATGCTGTAGGTTATTATGGTTTATTACAAGCATTTGCGGATGCTTTAAAATTATTATTAA
N  A  V  G  Y  Y  G  L  L  Q  A  F  A  D  A  L  K  L  L  L
*     *  *     *  *  *  *  *     *  *  *  *     *  *
                                                         300
AGGAATATATAGCGCCAACTCAAGCTAATATATTATTATTCTTTCTTGGTCCTATAATAA
K  E  Y  I  A  P  T  Q  A  N  I  L  L  F  F  L  G  P  I  I
*  *           *              *     *              *

CTTTAATTTTCTCTTTATTAGGTTATCTTGTTATTCCATTTGGTTCAGGTTTAGATATAT
T  L  I  F  S  L  L  G  Y  L  V  I  P  F  G  S  G  L  D  I
   *           *  *                 *              *
                                         400
CTGATTATAATTTAGGTATATTATATATGTTAGCTGTTAGTTCTTTAGCTACATATGGTA
S  D  Y  N  L  G  I  L  Y  M  L  A  V  S  S  L  A  T  Y  G
         *  *  *     *        *  *     *  *  *  *     *

TATTATTAGCTGGTTGATCAGCTAATTCTAAATATGCATTTTTAGGTTCATTAAGAAGTA
I  L  L  A  G  W  S  A  N  S  K  Y  A  F  L  G  S  L  R  S
*  *        *  *        *  *     *  *        *     *  *
                  500
CAGCTCAATTAATTAGTTATGAATTAATTTTAAGTTCTGTAATTTTATTAGTAATATTAT
T  A  Q  L  I  S  Y  E  L  I  L  S  S  V  I  L  L  V  I  L
   *  *     *  *  *  *        *              *           *
                                                         600
TAACAGGAAGTTTAAATATTATTACTATAATAGAATCACAAAGAGTAGTAAATTTATTAT
L  T  G  S  L  N  I  I  T  I  I  E  S  Q  R  V  V  N  L  L
      *  *     *        *     *        *              *  *

TCCCTTTATTCCCTTTATTCTTAGTATTCTTTATAGGTTCTATAGCAGAAACTAATAGAG
F  P  L  F  P  L  F  L  V  F  F  I  G  S  I  A  E  T  N  R
   *        *  *              *  *           *  *  *  *  *
                                          700
CTCCTTTTGATTTAGCAGAGGCAGAATCAGAACTTGTAAGTGGTTTCATGACAGAACATT
A  P  F  D  L  A  E  A  E  S  E  L  V  S  G  F  M  T  E  H
   *  *  *  *  *  *     *  *  *  *  *  *  *  *        *

CAGCTAGTATATTTGTATTCTTTTTCTTAGCTGAATATGCTAGTATTGTTTTAATTTGTA
S  A  S  I  F  V  F  F  L  A  E  Y  A  S  I  V  L  I  C
   *        *        *  *     *  *  *        *
                 800
TTTTAAATAGTATCTTATTTTTAGGTGGTTATTTAAGTATTATACCTTTAGATTTTATTG
I  L  N  S  I  L  F  L  G  G  Y  L  S  I  I  P  L  D  F  I
   *              *  *  *
                                                         900
TAAATATCTTAAATTTATTTGTTATTACTAATAATACTATTTTATTTGATATATTTGTAC
V  N  I  L  N  L  F  V  I  T  N  N  T  I  L  F  D  I  F  V

ATATTTCTTCTTCTCCTATTAATTTAGCATTAAAAACAGCTTTCTTAATCTTTGTATTCA
H  I  S  S  S  P  I  N  L  A  L  K  T  A  F  L  I  F  V  F
                                 *  *        *           *
                                          1000
TTTGAGTTAGAGCTTCTTTCCCTAGAATACGTTTTGATCAATTAATGTCTGTTTGTTGAA
I  W  V  R  A  S  F  P  R  I  R  F  D  Q  L  M  S  V  C  W
   *     *           *  *     *     *  *  *  *           *
```

Fig. 3(i).

```
CTGTTTTATTACCTATAATTATTGCTTATGTTGTATTAATACCTTGTATTGTATATGGTC
T  V  L  L  P  I  I  I  A  Y  V  V  L  I  P  C  I  V  Y  G
         *  *           *

                         1100
TAGGTATAATACCAACAAACATTTCATTATTATAGTAAAAAAAATAAAATTATATTATAA
L  G  I  I  P  T  N  I  S  L  L ***
         *  *

                                                           1200
ATAGTATAATTTTTAAATTATATGCGTTTTTAATAAAGGATTAGGGTACTATCCTATATA

TATTTTAGGAGTATTAAGTTAATTAATTATTAATATTAGTTATACTTAATATAATTGGCA
                                             1300
TACAATAAGTAAAAAAAAATAAATGATATATTACATTATCTAATGTAATATATATATTTT

ATAAATAGTAAATAAATTATTTAAATTATATATATATAATTATATGTCTTTATTATTATT
                                    START URF4  M  S  L  L  L  L
                         1400
AATAACAACTTTAATAGGATTACATTTAGTAACATTACAAGGTAATTATGGTTTATCCAT
 I  T  T  L  I  G  L  H  L  V  T  L  Q  G  N  Y  G  L  S  I
                                                     *
                                                           1500
AATTAATAATGTTAAAATAAAATCAATTGCGTTATTAACAACAATAATAAATTTGATTAT
 I  N  N  V  K  I  K  S  I  A  L  L  T  T  I  I  N  L  I  I
             *           *           *  *           *  *  *

ATCACTGGTAATGTTTATCTTATTTGATTTTAGTAGTAAACAATACCAATTTATAGAAGA
 S  L  V  M  F  I  L  F  D  F  S  S  K  Q  Y  Q  F  I  E  E
 *                    *
                                             1600
ACATTATGAAATTAATCATTTTGATATCTATTTAGGAGTAGATGGTTTATCAATATATTT
 H  Y  E  I  N  H  F  D  I  Y  L  G  V  D  G  L  S  I  Y  F
                                       *     *

TGTGTTATTAACAACAATAATAATGCCAATAGCTATATTATCTAATTGAAATTCAATAGA
 V  L  L  T  T  I  I  M  P  I  A  I  L  S  N  W  N  S  I  E
       *  *  *           *     *     *           *     *
                         1700
ATCTAAAAATGTATTATCATTTATAGTAATAATGCTATTGTTAGAAACACTTTTATTAGC
 S  K  N  V  L  S  F  I  V  I  M  L  L  E  T  L  L  L  A
                                *  *     *           *
                                                           1800
AGTGTTCTTAGTATTAGATATACTATTGTTTTACATCTTTTTTGAGAGTATATTACCACC
 V  F  L  V  L  D  I  L  L  F  Y  I  F  F  E  S  I  L  P  P
    *                          *  *  *  *  *  *        *     *

ATTATTTTTGTTAATAGGATTATTTGGTTCAAGTAATAAAGTAAGAGCTAGTTTTTATTT
 L  F  L  L  I  G  L  F  G  S  S  N  K  V  R  A  S  F  Y  L
             *           *                    *     *
                                             1900
ATTTTTATATACATTATTAGGATCATTATTTATGTTATTATCAATAATAGCTATTACTTC
 F  L  Y  T  L  L  G  S  L  F  M  L  L  S  I  I  A  I  T  S
       *  *  *     *  *  *        *           *

TATTATGGGTACATCAGATTTTGATGCATTAACAAAAGCAAACTTTAATTATATAACACA
 I  M  G  T  S  D  F  D  A  L  T  K  A  N  F  N  Y  I  T  Q
       *                       *  *
                         2000
AATATTTTTATTTTATGGTATATTTATAGCTTTTCGTGTAAAAACACCAGTAATGTTTTT
 I  F  L  F  Y  G  I  F  I  A  F  R  V  K  T  P  V  M  F  L
                                *  *     *  *                 *
                                                           2100
AAATACTTGATTATTAAAAGCTCACGTTGAATCACCTTTATCAGGAAGTATTATTTTAGC
 N  T  W  L  L  K  A  H  V  E  S  P  L  S  G  S  I  I  L  A
       *  *     *  *  *  *  *     *           *  *        *  *
```

Fig. 3(ii).

```
TGGTATAGTTTTAAAATTAAGTTTATACGGTATATTTAGATTAATTTTACCTTTATTACC
 G  I  V  L  K  L  S  L  Y  G  I  F  R  L  I  L  P  L  L  P
          *  *  *        *  *        *  *     *     *     *

                                      2200
TAAAGCTTCTATAAATTATACTTATATAATTTATGTTATAGGTGTAATAACTATATTATA
 K  A  S  I  N  Y  T  Y  I  I  Y  V  I  G  V  I  T  I  L  Y
                      *           *

TGCTAGTTTTAGTACATTAAGAACTATAGATATTAAAGAACTTATTGCTTATTCATCTGT
 A  S  F  S  T  L  R  T  I  D  I  K  E  L  I  A  Y  S  S  V
    *           *  *        *     *     *  *_ *  *  *  *
                   2300
ATCTCATGCAGCTGTATATTTAATAGGTGCATTTAGTAATACTATACAAGGTATTGAAGG
 S  H  A  A  V  Y  L  I  G  A  F  S  N  T  I  Q  G  I  E  G
 *  *     *                             *                 *
                                                        2400
ATCAATTGCTTTAGGTTTAGCTCACGGTTTTGTTTCTTCAGGTTTATTTATTTGTGCTGG
 S  I  A  L  G  L  A  H  G  F  V  S  S  G  L  F  I  C  A  G
          *        *  *  *        *  *     *  *        *

TGGTATCTTATACGATAGATCATCTACTAGATTAATAACTTATTATAGAGGTATGGCTCA
 G  I  L  Y  D  R  S  S  T  R  L  I  T  Y  Y  R  G  M  A  Q
          *     *           *                    *
                                      2500
AATTATGCCTATTTTCTCTGTGTTATTCTTCATATTAGCATTAGGTAATAGTGGAACTCC
 I  M  P  I  F  S  V  L  F  F  I  L  A  L  G  N  S  G  T  P
       *                                *     *           *

TTTAACTTTAAATTTCATAGGTGAGTTTATGTCATTATATGGAGTATTTGAAAGAATGCC
 L  T  L  N  F  I  G  E  F  M  S  L  Y  G  V  F  E  R  M  P
    *     *        *  *           *           *
                   2600
TATCTTAGGTGTTTTAGCTAGTACTTCTATAGTTTTCTCTGCTGCTTATACTATATTTAT
 I  L  G  V  L  A  S  T  S  I  V  F  S  A  A  Y  T  I  F  M
 *           *                          *     *           *
                                                        2700
GTATAATAGAATAGTATTTGGTGGTTCATATTCTATCTATTTTAGAGAAAATATAGGTGA
 Y  N  R  I  V  F  G  G  S  Y  S  I  Y  F  R  E  N  I  G  D
                   *

TGTAACTAGAAGAGAATTTATAATGTTATTAGTTTTCGTTATATTAACTGTATTATTTGG
 V  T  R  R  E  F  I  M  L  L  V  F  V  I  L  T  V  L  F  G
                                        *  *
                                       2800
TATATACCCTGCTCCTATTTTAGATGGTTTACATTATTCAGTTTCTTATTTAATATATAA
 I  Y  P  A  P  I  L  D  G  L  H  Y  S  V  S  Y  L  I  Y  N
                                  *  *

TATTAATTAA
 I  N ***
```

Fig. 3(iii).

Fig. 3. The sequence is that of the non-transcribed (sense) strand of *Bgl*II fragment 3 from a *Taq*I site upstream of the URF1 initiation codon to the URF4 termination codon. The amino acid sequences of URF1 and URF4 are indicated by the one-letter code (IUB-IUPAC, 1969), with those residues conserved between *A. nidulans* and *H. sapiens* marked by asterisks. There are no alternative reading frames to the two shown: URF1 contains 55 and 17 stop codons in the other two direct phases, and 61, 8 and 30 stop codons in the three inverse phases, while URF4 is closed 31 and 74 times in the direct phases and 73, 18 and 48 times in the inverse phases. The 42 codon segment of URF1 that displays no amino acid conservation with the human URF1 is underlined (see text).

AT content being more characteristic of the spacer regions between genes and the region of the LrRNA gene identified as a 'mini-insert' (Netzker *et al.*, 1982). Although an open reading frame through the whole of this region is evidence against the segment being removed during transcript processing, there are certain similarities between this 126-bp sequence and introns present in the *A. nidulans* genes for apocytochrome b (Waring *et al.*, 1982) and cytochrome oxidase subunit 1 (Waring *et al.*, in preparation). In particular, the consensus sequence for an upstream splice point in these introns is GGT↓NNN, the arrow indicating the cut site, (Davies *et al.*, in preparation) which suggests that a putative intron in *A. nidulans* URF1 could be spliced either between nucleotides 806 and 807 (GGT↓GGT) or between 809 and 810 (GGT↓TAT). These possible splice sites coincide exactly with the breakdown point for homology between the *A. nidulans*

URF1:

```
ASPERGILLUS:  MLYISIIEGLLVIVPALLSVAFVTVAERKTMASMQRRLGPNAVGYYGLLQAFADALKLLLKEYIAPTQANILLFFLGP
HUMAN:           MPMANLLLLIVPILIAMAFLMLTERKILGYMQLRKGPNVVGPYGLLQPFADAMKLFTKEPLKPATSTITLYITAP
BOVINE:          MFMINILMLIIPILLAVAFLTLVERKVLGYMQLRKGPNVVGPYGLLQPIADAIKLFIKEPLRPATSSASMFILAP
MOUSE:              INILTLLVPILIAMAFLTLVERKILGYMQLRKGPNIVGPYGILQPFADAMKLFMKEPMRPLTTSMSLFIIAP
                    *  ** *** ** *** *  ***       ** * *** ** ***** **** **  **   *     * ** * *
                    *  *  *   *  *      ***       ** * *** ** ** **  *** **  **   *            *

DROSOPHILA:       ILSLIGSLLLIICVLVSVAFLTLLERKVLGYIQIRKGPNKVGLMGIPQPFCDAIKLFTKEQTYPLLSNYLRYYISP
                       *      *   **      ***       * * *** **  *   *     ** **  **   *       *

IITLIFSLLGYLVIPFGSGLDISDYNLGILYMLAVSSLATYGILLAGWSANSKYAFLGSLRSTAQLISYELILSSVILLVILLTGSLNIITII
TLALTIALLLWTPLPMPNPL-VNL-NLGLLFILATSSLAVYSILWSGWASNSNYALIGALRAVAQTISYEVTLAIILLSTLLMSGSFNLSTLI
IMALGLALTMWIPLPMPYPL-INM-NLGVLFMLAMSSLAVYSILWSGWASNSKYALIGALRAVAQTISYEVTLAIILLSVLLMSGSFTLSTLI
TLSLTLALSLWVPLPMPHPL-INL-NLGILFILATSSLSVYSILWSGWASNSKYSLFGALRAVAQTISYEVTMAIILLSVLLMNGSYSLQTLI
*  *  **       *     *  *    ***** *** **** * *** **  *****   * **  ** ****  *      *  *  *  **  *  * *
   *  *        *     *         *** *  ** ***  ** **    ****   * **  ** ****         *     *  **     * *

IFSLFLSLFVWMCMPFFVKLYSF--NLGGLFFLCCTRLGVYTVMVAGW....
 *     *      *      *    *** *   *       *   *       **

ESQRVVNLLFPLFPLFLVFFIGSIAETNRAPFDLAEAESELVSGFMTEHSASIFVFFFLAEYASIVLICILNSILFLGGYLSIIPLDFIVNIL
TTQEHLWLLLPSWPLAMMWFISTLAETNRTPFDLAEGESELVSGFNIEYAAGPFALFFMAEYTNIIMMNTLTTTIFLGTTYDALSPELYTTYF
TTQEQMWLILPAWPLAMMWFISTLAETNRAPFDLTEGESELVSGFNVEYAAGPFALFFMAEYANIIMMNIFTTAILFLGTSHNPHMPELYTINF
TTQEHMWLLLPAWPMAMMWFISTLAETNRAPFDLTEGESELVSGFNVEYAAGPFALFFMAEYTNIILMNALTTIIFLGPLYYINLPELYSTNF
   *     ** *  **     **    ************ ********  *   *   *   ** **** *  *   **   *****
   *     ** *  *      **    ***** **** * ********  *   *   *   ** ***  *             ***

NLFVITNNTILFDIFVHISSSPINLALKTAFLIFVFIWVRASFPRIRFDQLMSVCWTVLLPIIIAYVVLIPCIVYGLGIIPTNISLL
VT-------------------------KTLLLTSLFLWIRTAYPRFRYDQLMHLLWKNFLPLTLALLMWYVSMPITISSIPPQT
TI-------------------------KSLLLTMSFLWIRASYPRFRYDQLMHLLWKNFLPLTLALCMWHVSLPILTSGIPPQT
MM-------------------------EALLLSSTFLWIRASYPRFRYDQLMHLLWKNFLPLTLALCMWHISLPIFTAGVPPYM
                           **  *   * * *** ** * ****    *    **    *                  **
                               *   * * *   ** * ****    *    **    *                   *
```

URF4:

```
ASPERGILLUS:  MSLLLLITTLIGLHLVTLQGNYGLSIINNVKIKSIALLTTIINLIISLVMFILFDFSSKQYQFIEEHYEINHFDIYLG
HUMAN:                          MLKLIVPTIMLLPLTWLSKKHMIWINTTTHSLIISIIPLLFFNQINNNLFS---------CSPTFS
BOVINE:                         MLKYIIPTIMLMPLTWLSKNNMIWVNSTAHSLLISFTSLLLMNQFGDNSLN---------FSLLFF
MOUSE:                          MLKIILPSLMLLPLTWLSSPKKTWTNVTSYSFLISLTSLTLLWQTDENYKN---------FSNMFS
                                   *      *        * * *    **    *****     **      *           *
                                                               *     **

VDGLSIYFVLLTTIIMPIAIL-S-NWNSIESKNVLSFIVIMLL-LETLLLAVFLVLDILLFYIFFESILPPLFLLIGLFGSS-NKVRASFYLF
SDPLTTPLLMLTTWLLPLTIMASQRHLSSEPLSRKKLYLSMLISLQISLIMTFTATELIMFYIFFETTLIPTLAIITRWGNQPERLNAGTYFL
SDSLSTPLLILTMWLLPLMLMASQHHLSKENLTRKKLFITMLISLQLFLIMTFTAMELILFYILFEATLVPTLIIITRWGNQTERLNAGLYFL
SDPLSTPLIILTAWLLPLMLMASQNHLKKDNNVLQKLYISMLISLQILLIMTFSATELIMFYILFEATLIPTLIIITRWGNQTERLNAGIYFL
 * **    ***  *  *  * *  *  *               **  *  **  *  *      ******* * *     *   *       * *
 * *     **   *  *   *  *                  **   *  *   *  *      *** **  * *     *   *       * *

LYTLLGSLFMLLSIIAITSIMGTSDFDALTKANF---NYITQIFLFYGIFIAFRVKTPVMFLNTWLLKAHVESPLSGSIILAGIVLKLSLYGI
FYTLVGSLPLLIALIYTHNTLGSLNILLLTLTAQELSNSWANNLMWLAYTMAFMVKMPLYGLHLWLPKAHVEAPIAGSMVLAAVLLKLGGYGM
FYTLAGSLPLLVALIYIQNTVGSLNFLMLQYWVQPVHNSWSNVFMWLACMMAFMVKMPLYGLHLWLPKAHVEAPIAGSMVLAAVLLKLGGYGM
FYTLIGSIPLLIALILIQNHVGTLNLMILSFTTHTLDASWSNNLLWLACMMAFLIKMPLYGVHLWLPKAHVEAPIAGSMILAAILLKLGSYGM
 *** ***  *   *  *    ** *  **           *     **          ** ** *   *  ** ***** *  ** *** * *** **
 *** **   *   *  *    *   *  **                              ** *  *      ** ***** *  **  **  ***  **

FRLILPLLPKASINYTYIIYVIGVITILYASFSTLRTIDIKELIAYSSVSHAAVYLIGAFSNTIQGIEGSIALGLAHGFVSSGLFICAGGILY
MRLTLILNPLT-KHMAYPFLVLSLWGMIMTSSICLRQTDLKSLIAYSSISHMALVVTAILIQTPWSFTGAVILMIAHGLTSSLLFCLANS-NY
LRITLILNPMT-DFMAYPFIMLSLWGMIMTSSICLRQTDLKSLIAYSSVSHMALVIVAILIQTPWSYMGATALMIAHGLTSSMLFCLANS-NY
IRISIILDPLT-KYMAYPFILLSLWGMIMTSSICLRQTDLKSLIAYSSVSHMALVIASIMIQTPWSFMGATMLMIAHGLTSSLLFCLANS-NY
 ** * * *          *   *        *  **  * * ********** *             *     *  **  *** ** **  *    *
 *  * * *          *            *  **  * * ****** ** *             *     *  *   *** ** **  *    *

DRSSTRLITYYRGMAQIMPIFSVLFFILALGNSGTPLTLNFIGEFMSLYGVFERMPILGVLASTSIVFSAAYTIFMYNRIVFGGSYSIYFREN
ERTHSRIMILSQGLQTLLPLMAFWWLLASLANLALPPTINLLGELSVLVTTFSWSNITLLLTGLNMLVTALYSLYMFTTTQWGSLTHHINNMK
ERIHSRTMILARGLQTLLPLMATWWLLASLTNLALPPTINLIGELFVVMSTFSWSNITIILMGVNMVITALYSLYMLIMTQRGKYTYHINNIS
ERIHSRTMIMARGLQMVFPLMATWWLMASLANLALPPSINLMGELFITMSLFSWSNFTIILMGINIIITGMYSMYMIITTQRGKLTNHMINLQ
 *    *       **      *           * *   * * * ***   *    *       *    *     ** * *    *       *
 *    *       *       *           * *   * * * **    *    *       *    *        *      *       *

IGDVTRREFIMLLVFVILTVLFGIYPAPILDG  LHYSVSYLIYNIN
PSFTRENTLMFMHLSPILLLSLNPDIITGFSS  CKYSLTKTSDCESDN
PSFTRENALMSLHILPLLLLTLNPKIILGPLY  CKYSLTKTLDCESNN
PSHTRELTLMALHMIPLILLTTSPKLITGLTM  CEYSLQKTLDCESDNSK
        *     **             *     **
                                   **
```

Fig. 4. Comparison between the derived amino acid sequences of URF1 and URF4 from the *A. nidulans*, human (Anderson *et al.*, 1981), bovine (Anderson *et al.*, 1982) and mouse (Bibb *et al.*, 1981) mitochondrial genomes. The sequences are written in the one-letter code (IUB-IUPAC, 1969). Asterisks indicate conserved residues, with homology determined on two levels. The upper row of asterisks marks those amino acids of the *A. nidulans* sequence that are present in the same position in at least one of the three mammalian species. The lower row of asterisks shows those residues that are conserved between all four species. The partial sequence of *Drosophila melanogaster* URF1 (Clary *et al.*, 1982) is also shown, with asterisks marking residues conserved in all five translation products. The last 13 amino acids of *A. nidulans* URF4 are compared with the extended mammalian polypeptides (see text).

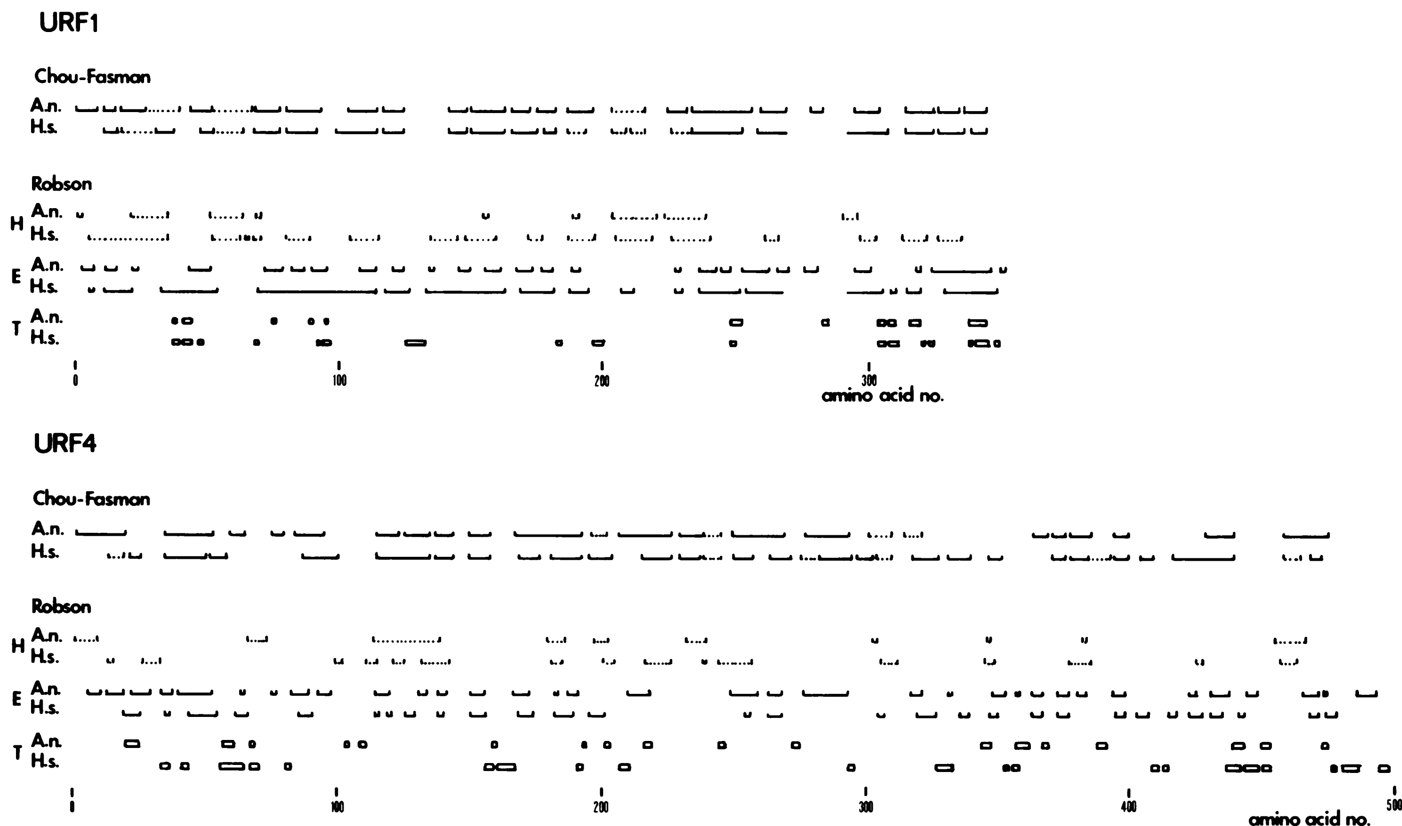

Fig. 5. Secondary structure predictions for the URF1 and URF4 gene products. Using the Chou and Fasman method, possible α-helix and β-strand regions were identified and the published guidelines used to predict which structures are likely to nucleate and to determine the probable boundary points. Helices are shown as dotted lines and β-strands as unbroken lines. The Robson method was used to calculate the probability of each residue participating in each of three different conformations. The figure indicates those regions with a high probability (unbiased information content >100 centinats) of forming either helices (dotted lines), extended chains (roughly equivalent to β-strands, shown by unbroken lines) and turns (open boxes). Exact predictions of the most probable conformation in each region were not made using the Robson technique because of the difficulty in choosing the correct decision constant without physical information about the proteins. Several regions are therefore shown as possessing high probability values for more than one conformation. The high overall β-structure of URF1 and URF4 suggests that, were a decision constant used, the amount of α-helix predicted would be less than that shown in the figure. Abbreviations: H, helix; E, extended chain; T, turn; A.n., *A. nidulans*; H.s., *H. sapiens*.

and mammalian products.

Primary nucleotide sequence conservation between the fungal and human genes is of the same order as the amino acid homology (41% for URF1, 36% for URF4). However, very few triplet codons are conserved in every nucleotide. In URF1 only 26 out of the 352 amino acids (7.4%) are coded by precisely the same triplet in both the *A. nidulans* and human genes; the equivalent figures for URF4 are 29 precisely conserved codons out of 488 (5.9%). To compensate there are 136 silent base pair changes in URF1 and 122 in URF4.

Further points concerning the URF primary sequences are worth noting. Firstly, of the nine possible protein-coding genes and three intron maturase reading frames thus far sequenced in the *A. nidulans* mitochondrial genome, only URF1 uses UAG as the termination codon. Although several human mitochondrial genes terminate with UAG (Anderson *et al.*, 1982), this codon is not found in either mouse (Bibb *et al.*, 1981) or bovine (Anderson *et al.*, 1982) genomes, and has not yet been encountered in yeast (Bonitz *et al.*, 1980).

A second point concerns the mammalian gene URF4L, a 97 or 98 codon reading frame that forms a short, out-of-phase overlap with the 5′ end of URF4 in the human, mouse and bovine genomes. RNA transcript mapping has indicated that URF4L is transcribed with URF4 into one polycistronic message (Montoya *et al.*, 1981; Ojala *et al.*, 1981), and it has been speculated that the two reading frames are spliced together prior to translation (Anderson *et al.*, 1981), although there is evidence that human mitochondria contain a mitochondrially-encoded protein of approximately the size expected for the unprocessed URF4L gene product (Bhat *et al.*, 1982). We have no evidence of URF4L in the *A. nidulans* mitochondrial genome. There is no reading frame upstream of URF4 that could be equivalent to the mammalian URF4L, neither does any part of the *A. nidulans* URF4 gene show recognisable homology with this reading frame. Although it is possible that URF4L is present in the small region of the *A. nidulans* genome yet to be characterised, it seems highly unlikely that URF4 itself undergoes any major rearrangement, or is joined to any additional reading frame, during processing in *A. nidulans*.

The final point concerns the 3′-terminal region of URF4. In mammals the possibility exists that extended polypeptides could be translated by read-through into adjacent downstream genes prior to completion of processing. Comparisons between the bovine and human sequences (Anderson *et al.*, 1982) have revealed that in the case of URF4, read-through into the histidine tRNA gene would result in a polypeptide 15 amino acids longer than the product of the mature polyadenylated message. The additional amino acids would be highly conserved between the two species, evidence that such

an extended translation product may in fact be synthesized. The *A. nidulans* URF4 gene product is 13 amino acids longer than the non-extended mammalian URF, and two contiguous amino acids are conserved when the extended mammalian URF products are compared with the *A. nidulans* sequence (Figure 4). The fungal URF could possibly be more equivalent to this extended read-through product of the mammalian URF4 genes than to the product of the processed mammalian message.

Conservation of secondary structures between fungal and human URFs

Secondary structure characterisation, following the system of Chou and Fasman (1978), indicates that both URF1 and URF4 are predominantly β-sheet proteins with only limited helical regions. Twenty-one individual β-strands and only 13 helices are predicted for *A. nidulans* URF1 and 20 β-strands plus four helices for URF4 (Figure 5). Analyses by the method developed by B. Robson and collaborators (Garnier *et al.*, 1978) are in broad agreement, with large regions of each URF predicted as possessing a high probability for forming extended chains with rather fewer likely helical structures (Figure 5). Comparisons between the predicted secondary structures of the fungal and human URFs reveals that, despite the differences between the primary sequences, the conformation of the gene products are highly similar. The predictions of Chou and Fasman show that five β-structures are precisely conserved in URF1, and another 10 β-strands and two helices are present in closely equivalent positions. Similarly, in URF4, one α-helix and three β-strands are exactly conserved between the fungal and human URFs, with an additional 13 strands and one helix occupying equivalent positions on the polypeptide chains. Moreover, most structures predicted by the Robson technique as possessing a high probability in the fungal URF coincide in position with similar predicted structures in the human products.

The Robson predictions indicate that some turns may be conserved between the fungal and human URFs. Although the data represent probabilities and not definite structures, of the seven turns that are predicted in *A. nidulans* URF1, six are at equivalent positions to predicted turns in the human gene product. Similarly, in URF4, three out of the nine appear to be conserved.

In several cases a strong secondary structure prediction is associated with close conservation of amino acid sequence between the *A. nidulans*, human, bovine and mouse gene products. For example, the best-conserved region of URF1, a negatively-charged stretch of 25 amino acids (AET....SGF) coincides with a strong α-helix predicted by both Chou-Fasman and Robson methods. Similarly, the conserved sequences KAHVE and LIAYSSVSH in URF4 could also form strong helices. The implication that these helical regions are involved in important functional activities cannot be tested at present.

In several stretches of both URF1 and URF4 the predicted conformation of the *A. nidulans* and *Homo sapiens* products are similar even though the sequence homology is relatively low. An example is the region of URF1 between amino acids 150 and 195 (Figure 5) where the amino acid homology is 35%, somewhat less than the overall figure for the gene.

The intergenic region

The genes of URF1 and URF4 are separated in the *A. nidulans* genome by a 247-bp spacer region in which the AT content rises to 89%, compared with values of 75% for URF1 and 76% for URF4. The nucleotide sequence of the spacer region is such that base pairing within a single strand could produce a series of loop structures one of which is similar in shape, though not sequence, to a base paired loop present in a region shown to contain a replication origin in the *A. amstelodami* mitochondrial genome (Lazarus and Kuntzel, 1981). An alternative to replication origins is that one or other of the possible loops function as transcription recognition signals involved in processing of the URF genes. A full analysis of these intergenic loop structures awaits completion of the genome sequence.

Discussion

At least eight unidentified reading frames are present in the *A. nidulans* mitochondrial genome. Six of these are not homologous with any of the human URFs and have, as yet, no known counterparts in other lower organisms. These six genes are: URF A1, a 228-residue reading frame located between the 16S rRNA and the structural gene for cytochrome oxidase subunit 3 (Netzker *et al.*, 1982); the short (27 residues), highly polar URF A2, positioned upstream of the histidine tRNA (Netzker *et al.*, 1982); three reading frames clustered together upstream of the *cob*A locus (unpublished data); and the 48 codon URFx immediately upstream of the gene for ATPase subunit 6 (Grisi *et al.*, 1982). In contrast, the two additional unidentified reading frames in *A. nidulans* display clear homology with human URFs. The equivalence of these genes with the human URFs 1 and 4 is apparent from the derived primary amino acid sequences and is further emphasised by protein secondary structure comparisons. Predictions using both the Chou-Fasman and Robson techniques suggest that, within both URF1 and URF4, long stretches of the gene product sequences are strongly conserved at the secondary structure level, with recognisable similarities at most other points in the chains. This structural homology is important as it is a clear indication that functional translation products of URF1 and URF4 are synthesised in fungal and mammalian mitochondria.

When the human mitochondrial DNA sequence was published (Anderson *et al.*, 1981) it seemed probable that the URFs were exclusively mammalian genes with no counterparts in lower organisms. These opinions have to be modified now that homologues of URF1 and URF4 have been discovered in a fungal mitochondrial genome. The presence of these genes in such unrelated species as *A. nidulans* and *H. sapiens* is a clear indication that at least URF1 and URF4 are of general functional importance and are likely to be found in either the mitochondrial or nuclear genomes of other organisms. Exploiting *A. nidulans*, we can now attempt to identify the functions of two URF gene products.

Materials and methods

DNA preparations

The strain used was *A. nidulans yA2, pyroA4, cnxC3*. Techniques for the preparation of mitochondrial DNA, replicative forms of M13 vectors and clones, single-stranded DNA of M13 clones, DNA restriction fragments, sequencing primers and M13 clones in reversed orientations, have all been previously described (Waring *et al.*, 1981).

Cloning and DNA sequencing

Templates for sequencing the region containing URF1 and URF4 were prepared as clones of total mitochondrial DNA cut with *Bgl*II and/or *Bcl*I and cloned into the *Bam*HI site of either M13mp6 or pBR322. *Bgl*II fragment

3 (for nomenclature see Waring *et al.*, 1981) was restricted with *Sau*3A, *Alu*I or *Taq*I and subcloned into the *Bam*HI, *Hinc*II or *Acc*I sites of M13mp7 (Messing *et al.*, 1981). DNA sequencing was performed by the chain termination method (Sanger *et al.*, 1977) using either the 30-bp universal primer (Anderson *et al.*, 1980) or internal primers prepared from specific restriction fragments (Waring *et al.*, 1981).

Protein secondary structure predictions

Two methods were used to predict protein secondary structures. The technique of Chou and Fasman (1978) was employed to manually locate possible α-helix and β-strand regions within primary sequences. In addition, the method developed by Robson and collaborators (Garnier *et al.*, 1978) was used in conjunction with a mini-computer to calculate information measures for helix, extended chain, turn and random coil.

Acknowledgements

We wish to thank B. Robson for advice and computer facilities during the secondary structure predictions. The work was supported by a project grant from the Medical Research Council to R.W.D and C.S.

References

Anderson,S., Gait,M.J., Mayol,L. and Young,I.G. (1980) *Nucleic Acids Res.,* **8**, 1731-1743.

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.

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.

Bhat,N.K., Niranjan,B.G. and Avadhani,N.G. (1982) *Biochemistry (Wash.),* **21**, 2452-2460.

Bibb,M.J., Van Etten,R.A., Wright,C.T., Walberg,W.M. and Clayton,D.A. (1981) *Cell,* **26**, 167-180.

Bonitz,S.G., Coruzzi,G., Thalenfield,B.E., Tzagoloff,A. and Macino,G. (1980) *J. Biol. Chem.,* **255**, 11927-11941

Chou,P.Y. and Fasman,G.D. (1978) *Annu. Rev. Biochem.,* **47**, 251-276.

Clary,D.O., Goddard,J.M., Martin,S.C., Fauron,C.M.-R. and Wolstenholme,D.R. (1982) *Nucleic Acids Res.,* **10**, 6619-6637.

Davies,R.W., Scazzocchio,C., Waring,R.B., Lee,S., Grisi,E., Berks,M.M. and Brown,T.A. (1982) in Slonimski,P., Borst,P. and Attardi,G. (eds.), *Mitochondrial Genes,*, Cold Spring Harbor Laboratory Press, NY, pp. 405-410.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.,* **120**, 97-120

Grisi,E., Brown,T.A., Waring,R.B., Scazzocchio,C. and Davies,R.W. (1982) *Nucleic Acids Res.,* **10**, 3531-3539.

Köchel,H.G. and Küntzel,H. (1981) *Nucleic Acids Res.,* **9**, 5689-5696.

Köchel,H.G., Lazarus,C.M., Basak,N. and Küntzel,H. (1981) *Cell,* **23**, 625-633.

Lazarus,C.M. and Küntzel,H. (1981) *Curr. Genet.,* **4**, 99-107.

Messing,J., Crea,R. and Seeburg,P.H. (1981) *Nucleic Acids Res.,* **9**, 309-321.

Montoya,J., Ojala,D. and Attardi,G. (1981) *Nature,* **290**, 465-470.

Netzker,R., Köchel,H.G., Basak,N. and Küntzel,H. (1982) *Nucleic Acids Res.,* **10**, 4783-4794.

Ojala,D., Montoya,J. and Attardi,G. (1981) *Nature,* **290**, 470-474.

Sanger,F., Nicklen,S. and Coulson,A.R. (1977) *Proc. Natl. Acad. Sci. USA,* **74**, 5463-5467.

Tzagoloff,A., Macino,G. and Sebald,W. (1979) *Annu. Rev. Biochem.,* **48**, 419-441.

Waring,R.B., Davies,R.W., Lee,S., Grisi,E., Berks,M.M. and Scazzocchio, C. (1981) *Cell,* **27**, 4-11.

Waring,R.B., Davies,R.W., Scazzocchio,C. and Brown,T.A. (1982) *Proc. Natl. Acad. Sci. USA,* **79**, 6332-6336.

Waring,R.B. and Scazzocchio,C. (1982) *Genetics,* in press.

Note added in proof

Recent sequencing of the *A. nidulans* genome in the region downstream of *oxi*B has revealed a possible homologue of the mammalian URF5.