Structural analysis of the sequence coding for an inducible 26-kDa protein in human fibroblasts

Guy HAEGEMAN, Jean CONTENT, Guido VOLCKAERT, Rik DERYNCK, Jan TAVERNIER and Walter FIERS

Laboratory of Molecular Biology, State University of Ghent

When human fibroblast cells were stimulated with poly(I) · poly(C) in the presence of cycloheximide for the production of interferon-β (IFN-β), a 26-kDa protein could be immunoprecipitated by antiserum raised against partially purified human IFN-β [Content, J., De Wit, L., Pierard, D., Derynck, R., De Clercq, E. & Fiers, W. (1982) *Proc. Natl Acad. Sci. USA 79*, 2768−2772]. In our hands this 26-kDa protein showed no antiviral activity. Other investigators have, however, reported the presence in the same conditions of a second type of IFN, a so-called β_2 species [Weissenbach, J., Chernajovsky, Y., Zeevi, M., Shulman, L., Soreq, H., Nir, U., Wallach, D., Perricaudet, M., Tiollais, P. & Revel, M. (1980) *Proc. Natl Acad. Sci. USA 77*, 7152−7156] of which the mRNA structure and protein characteristics strongly suggests identity with the 26-kDa product.

In this paper we describe the nucleotide sequence of the 26-kDa cDNA and part of the corresponding genomic clone. The cDNA clones were isolated from a library made with mRNA from induced human fibroblasts. As, however, the information thus obtained was still incomplete, genomic clones were isolated from a total human DNA library. In this way, the entire region coding for the 26-kDa protein was established, as well as the neighbouring sequences including the inducible promoter area. From the deduced polypeptide sequence a number of characteristics of the 26-kDa protein can be explained. It turns out that the 26-kDa protein gene and the so-called 'IFN-β_2' gene are identical. However, extensive homology searches indicate that the 26-kDa protein does not show statistically significant sequence homology with any known interferon species. Hence, the question of whether the 26-kDa product represents a novel IFN species remains open.

Upon induction of human fibroblast cells to produce IFN-β by treatment with poly(I) · poly(C) in the presence of cycloheximide, synthesis of a variety of other proteins, ranging in size from 15 to 70 kDa, is also stimulated [1]. Weissenbach et al. [2] reported on an mRNA species of about 14S which, they believed, coded for a novel fibroblast-type IFN, named 'IFN-β_2' [2]. Whether one of the IFN mRNAs described by Sehgal et al. [3, 4] is related to the former is not known. We have also characterized a 14S mRNA species present upon induction of several different fibroblast cell lines; this 14S mRNA species can be translated *in vitro* to a 26-kDa protein [5]. The latter is immunoprecipitated by goat antiserum raised against partially purified human IFN-β, but not by goat antiserum raised against highly purified IFN-β. The structural relationship of the 26-kDa protein to 22-kDa and 27-kDa products secreted in the growth medium upon induction of fibroblast cells was clearly demonstrated [5]. However, although the 26-kDa protein and the so-called 'IFN-β_2' are undoubtedly identical (see Discussion), we were not able to detect biological IFN activity either after injection of the purified 14S mRNA fraction in *Xenopus* oocytes or by partial purification of the proteins secreted into the medium by the induced cells [5]. The fact that other investigators were able to show some IFN activity corresponding to the 14S mRNA translation product may possibly result from a more sensitive assay for biological IFN activity [2].

Induction of the 26-kDa protein is not at all affected by the presence of the double-stranded IFN inducer poly(I)·poly(C), but is in fact activated by the cycloheximide treatment of the cells [6]. This was a first indication that expression of the 26-kDa protein does not share the general regulatory characteristics of type I IFN [6]. In fact, a cycloheximide-dependent induction, in the absence of poly(I) · poly(C), was also reported for the production of the so-called 'IFN-β_2' [2].

Recently, it was noted that an equally powerful production of 26-kDa protein can be obtained following induction of fibroblast cells with a lymphokine, subsequently identified as interleukin 1 [7]. This stimulation is still further enhanced by the presence of cycloheximide [7].

In this paper, we report the entire coding information and the deduced polypeptide sequence of the 26-kDa protein, as well as the nucleotide sequence of the neighbouring untranslated regions and the inducible promoter area. The structure and organization of the corresponding chromosomal gene will be published elsewhere.

MATERIALS AND METHODS

mRNA isolation and cDNA cloning

Induction of the human diploid fibroblast cell line VGS, or of the human osteosarcoma cell line MG63, was performed as described previously [5, 8]. For the mRNA isolation and the cDNA cloning we followed the procedure described by Derynck et al. [8] and Devos et al. [9].

The use of crude anti-IFN-β antiserum to demonstrate the presence of a specific 26-kDa protein derived from cell-free

translation of total mRNA of induced cells or mRNA selected on 26-kDa-protein cDNA clones has been described by Content et al. [5].

Screening human genomic DNA and subcloning

A human gene library, constructed and kindly provided by Dr T. Maniatis [10], was screened for the presence of genomic 26-kDa-protein-specific nucleotide sequences by blotting on nitrocellulose filters according to Benton and Davis [11] followed by plaque hybridization *in situ*.

Further subcloning of relevant genomic *Eco*RI fragments was done in the plasmid vector pUR250, which allows a rapid screening for positive clones on indicator plates [12].

Probes

The probes used were prepared and labelled in different ways depending on their respective sizes. The internal 26-kDa-protein-specific *Sau*3A cDNA fragment of 480 bp (Fig. 1), used for the genomic screening, was isolated from a total 26-kDa-protein cDNA digest and purified on low-melting-point agarose. The purified cDNA fragment was labelled by the nick-translation procedure according to Rigby et al. [13]. Entire chimaeric cDNA plasmids, such as 26K-1 or 26K-2, were labelled in the same way.

Shorter cDNA fragments, such as the 163-bp *Hin*fl probe or the 200-bp *Rsa*I probe (Fig. 1), were labelled externally at their 5' ends by kination [14] on a dephosphorylated restriction digest of the appropriate total cDNA (i.e. 26K-2 or 26K-7). This reaction mixture was separated on a 5% polyacrylamide gel and the relevant fragment was recovered by elution with 1 M NaCl.

The single-stranded *Hin*fl cDNA primer was prepared by heating the solution (containing 40% dimethylsulfoxide and 1 mM EDTA) of the end-labelled double-stranded cDNA fragment at 37 °C for 5 min and rapid cooling to −70 °C. The mixture was then thawed at 0 °C and loaded on a 10% polyacrylamide gel (bisacrylamide/acrylamide = 1:50). Recovery of the separated single strands was again with 1 M NaCl in siliconized test tubes.

Sequence determination

The nucleotide-sequence determination of the cDNA inserts or of the genomic DNA segments was performed following the chemical modification and degradation procedure of Maxam and Gilbert [15]. Appropriate fragments were end-labelled, either by dephosphorylation and kination [14], or by filling-in at the 3' end with Klenow polymerase [16].

The reaction products were fractionated on thin gels (0.3 mm) of 6% or 8% polyacrylamide (bisacrylamide/acrylamide = 1:20).

Analytical reverse transcription

Total poly(A)-containing RNA from induced cells was fractionated by sucrose-gradient centrifugation, and 26-kDa-protein-specific sequences were localized by dot-blot hybridization. The RNA from the corresponding fractions was precipitated, redissolved in water, and to this a 10-fold molar excess of the end-labelled, complementary strand of the *Hin*fl cDNA fragment was added as a primer for reverse transcription. This mRNA-primer DNA solution was lyophilized and

the residue redissolved in 30 μl hybridization buffer (0.4 M NaCl; 40 mM Pipes buffer, pH 6.4; 1 mM EDTA). The mixture was kept at 80 °C for 5 min, allowed to reanneal for 3 h at 45 °C and precipitated with ethanol at −20 °C. Subsequently, the RNA · DNA hybrids were dissolved in reverse-transcription buffer essentially made up as described by Devos et al. [9], to which reverse transcriptase was added in a ratio of 5−10 units/μg RNA. The reaction was allowed to continue for 30 min at 41 °C, stopped with EDTA at a final concentration of 25 mM and extracted with a double volume of phenol/chloroform (1:1). Following precipitation, the reaction mixture was dissolved in a mixture of 90% formamide and 10% 0.1 M NaOH, heated for 3 min at 90 °C and applied on a 5% denaturing polyacrylamide gel for electrophoresis.

RESULTS

Sequence determination of isolated cDNA clones

The preparation and size-fractionation of mRNA from induced human fibroblast cells and the construction of the corresponding cDNA clones in the *Pst*I site of the plasmid pBR322 have been described previously [8]. Selection of a clone containing a 26-kDa-protein cDNA insert was based on hybridization, elution-translation [8], followed by immunoprecipitation with 'crude' anti-IFN-β antiserum and electrophoresis in a denaturing gel. This 'crude' antiserum has been described before [5] and allows immunoprecipitation of the 26-kDa protein as well as of IFN-β.

More 26-kDa-protein clones were subsequently picked up from the cDNA library by colony hybridization with a nick-translated DNA probe of the first clone, named 26K-1. A total of 16 cDNA clones was thus obtained, which contained insert sequences ranging in size between 720 and 850 bp (the dsDNA had been size-selected on polyacrylamide gel during the cloning procedure [8]). Several cDNA inserts start from the 3'-terminal poly(A) tail of the mRNA, except for one clone (26K-7) which displayed a different pattern upon restriction-enzyme analysis. The cDNA of this clone starts at an internal position in the 3'-untranslated region (presumably due to nicking of the DNA during the cloning procedure) and correspondingly extends about 110 bp further in the 5'-terminal direction (Fig. 1).

It was found after sequence analysis however, that the size of the cDNA inserts was not sufficient to encompass the total coding part of the protein, as a fairly large, 3'-untranslated region (>460 bp) is present. Hence, a search for 26-kDa-protein coding sequences extending further in the 5' direction was needed to deduce the structure of the entire 26-kDa polypeptide.

Sequence determination of the cDNA inserts was carried out according to the chemical degradation procedure of Maxam and Gilbert [15]. The sequencing strategy is summarized in Fig. 1. The different types of inserts overlap each other and together they form a continuous sequence. No inversions or sequence rearrangements have been observed, as occurred in the IFN-β clones of the same cDNA library [17]. One open reading frame runs throughout the whole nucleotide sequence, coding for about 180 amino acids. Considering the size of a cell-free translation product of 26 kDa, a polypeptide of approximately 210−220 amino acids would be expected. Hence, the sequence information for another 30−40 amino acids (100−120 bp) is missing from the collection of cDNA clones analyzed.
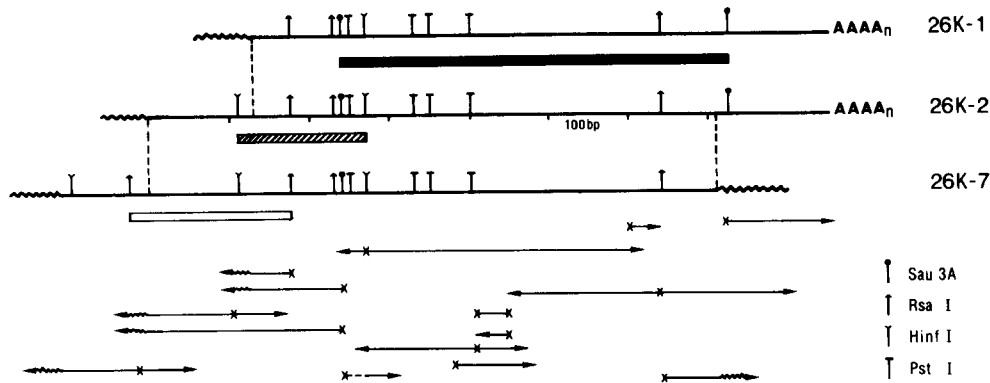
Fig. 1. *26-kDa-protein cDNA diagrams.* Three different 26-kDa-protein cDNA inserts (named 26K-1, 26K-2 and 26K-7) are represented by a straight line, the length of which is proportional to their respective size; $AAAA_n$ denotes the position of the 3'-terminal poly(A) tail; wavy lines represent surrounding pBR322 vector sequences; the vertical dashed lines position the individual cDNA inserts relative to each other. Relevant restriction sites are indicated. Underlying blocks show the position of defined restriction fragments of interest (see text), i.e. the internal 480-bp *Sau*3A fragment (in black), the 163-bp *Hin*fI fragment (hatched block) and the 200-bp *Rsa*I fragment (open bar). The horizontal arrows at the bottom indicate the sequencing strategy (a dashed section marks the connection between a labelled restriction site and the sequencing ladder which was not read from the very beginning)

Isolation of '26-kDa-protein' gene sequences from the human genomic library

As cDNA clones often do not contain the information up to the very first nucleotide of the mRNA, we preferred to switch from a cDNA cloning procedure to the use of genomic DNA as the starting material for the further extension of the nucleotide sequence. In addition, this would also allow us to gain some insight into the structure of the 26-kDa-protein gene and, what is possibly more interesting, to determine the putative promoter area preceding the mRNA information.

Consequently, a human gene library [10] was screened by plaque hybridization with the internal *Sau*3A fragment of 26-kDa-protein cDNA (Fig. 1), labelled with ^{32}P by nick-translation. In this way, four individual recombinant phages out of a total number of $6 - 7 \times 10^5$ were isolated. Three were apparently identical, while the fourth displayed a differing band pattern upon restriction-enzyme analysis. Addition of the chain lengths of all the restriction fragments of a given genomic insert gave a total of about 14 000 bp, which may correspond to the complete Charon 4A phage insert.

Remarkably, the two types of 26-kDa-protein phage displayed almost no identical fragments following digestion with a number of restriction enzymes, except for two defined bands obtained after *Hin*dIII degradation. To reconfirm the nature of the inserted genomic DNA segments, Southern blotting was performed on a number of restriction digests, followed by hybridization under stringent conditions with the most complete 26-kDa-protein cDNA information, i.e. 26K-2 (results not shown). As expected, we found positive signals for some of the restriction fragments in both phage types. This confirmed that the two types of phages are at least partly homologous with the cDNA probe and are presumably derived from the same 26-kDa-protein genomic gene; it further illustrates that this 26-kDa-protein coding information is dispersed over a fairly wide genomic area, covered here by the two almost non-overlapping phage inserts. Furthermore, we learned from this experiment which digestion fragments correspond to 26-kDa-protein cDNA sequences, although no identical 26-kDa-protein specific band was found in the comparable digests from the two genomic inserts.

One of the phage DNAs was used to probe the other type, and again the two above-mentioned bands in the *Hin*dIII

digest were shown to be identical. From that, we concluded that these bands are in fact common to the two types of genomic insert and do actually represent an overlapping region. They are derived from non-coding sequences (e.g. intron segments), as they were not detected by the 26-kDa-protein cDNA probe. Hence, both insert types are derived from the same genomic region and share a common gene segment.

When the digested phage inserts were probed with the internal *Hin*f fragment (Fig. 1), which, although slightly overlapping the *Sau*3A probe used previously, maps distinctly in the 5' portion of the cDNA, only one single band in the restriction pattern of one type of phage became labelled, while none of the bands in the digests of the second phage type hybridized. This confirmed our previous conclusion concerning the orientation of the phage inserts and at the same time demonstrated which of the phage inserts corresponds to the 5' end. The structural organization of this particular phage insert, as deduced from double restriction-digestion analysis, is partially represented in Fig. 2.

Subcloning genomic fragments and sequence analysis

In order to make the genomic sequences readily available for further sequence analysis, each of the *Eco*RI fragments of interest was subcloned in pUR250 and subjected to detailed restriction-enzyme analysis. One subclone, g26K-14, contained a 2200-bp *Eco*RI fragment, to which the previously mentioned *Hin*fI cDNA segment hybridized specifically. This *Eco*RI fragment contained a unique *Hin*dIII site and a unique *Pst*I site (Fig. 2). In order to verify whether the latter is identical with the unique *Pst*I site present in the *Hin*fI cDNA probe (Fig. 1), the genomic subclone was labelled at this *Pst*I position and partially sequenced. Reading into the 5' direction, the genomic sequence and the corresponding cDNA were identical over an area of at least 100 nucleotides, confirming our localization of these cDNA sequences on the chromosomal DNA. Nevertheless, at the 3' side of this *Pst*I site, the genomic sequence rapidly diverged from the cDNA and undoubtedly entered into an intron region. As, however, intron sequences may also be expected in the more upstream regions of the genomic DNA (i.e. at the 5' side of the characteristic *Pst*I site), the genomic subclone g26K-14 was further characterized by digestion with a variety of restriction
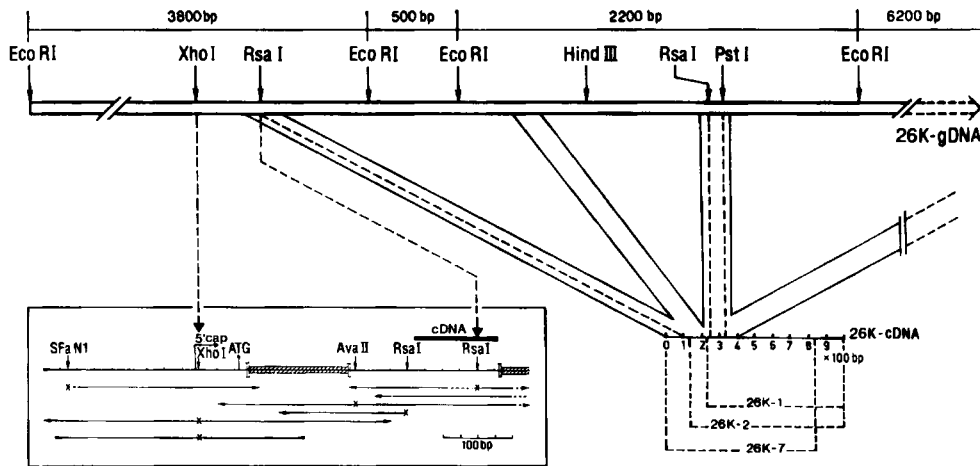
Fig. 2. *Schematic representation of the 5' portion of the chromosomal 26-kDa-protein gene.* The map shows part of a 26-kDa-protein genomic DNA (26K-gDNA) phage insert, carrying 5'-proximal sections of the 26-kDa-protein cDNA. The genomic *Eco*RI fragments studied, as well as their respective restriction sites of interest, are shown. On the same scale, the 26-kDa-protein cDNA (26K-cDNA) (incomplete) is represented and the distribution of 5'-proximal cDNA sequences over the genomic DNA is indicated. The insert figure displays a more detailed structural map, corresponding to the physical 5' end of the chromosomal 26-kDa-protein gene. Restriction sites relevant to structural analysis, as well as biologically interesting sites (such as the putative 5'-terminal cap site and the translational initiation codon ATG), are indicated. Hatched regions on the diagram line correspond to intervening sequences; the black box represents part of the cDNA and shows the overlap between the established cDNA sequence and the genomic DNA. Arrows underneath indicate the sequencing strategy (dashed sections represent parts which were not read)

enzymes. Following Southern blotting, it was tested again with another probe, namely the 200-bp *Rsa*I cDNA fragment, derived from clone 26K-7 (Fig. 1) which extends still further into the 5'-terminal direction than the *Hin*fI probe used previously. This experiment confirmed that part of these cDNA sequences were indeed present in the 2200-bp *Eco*RI fragment, but separated into two regions, following and preceding the *Hin*dIII site, thus demonstrating the presence of an intron in between (Fig. 2). Furthermore, the most 5'-extended portion of the probe had no counterpart in the 2200-bp fragment and is therefore to be localized still further upstream in the chromosomal DNA. Referring to the phage map (Fig. 2), the preceding genomic *Eco*RI fragment of 500 bp was tested and did not give any hybridization signal with the 200-bp *Rsa*I probe; the more 5'-oriented DNA segment of 3800 bp, however, did hybridize. Restriction-enzyme analysis of the latter fragment, followed by Southern blotting and probing with the 200-bp *Rsa*I fragment, finally showed the approximate localization of the remaining cDNA sequences in this newly identified genomic DNA fragment (Fig. 2). In this way, at least three introns had been localized as well.

An outline of the sequencing strategy, as carried out on the relevant part of the 3800-bp *Eco*RI fragment by the Maxam and Gilbert procedure [15], is presented in Fig. 2. An overlap of at least 100 bp between the established cDNA sequence and part of the genomic 3800-bp *Eco*RI fragment links the two DNA sequences. The nucleotide sequence of the DNA region of interest revealed a unique reading frame continuing the coding information of the cDNA into the 5' direction up to the putative initiation codon ATG. This triplet is preceded by a purine nucleotide at position -3 and furthermore by two C residues at -5 and -2, and thus fits in quite well with consensus rules for eukaryotic initiation sites [18]. Proceeding further upstream, we entered the 5'-untranslated mRNA region (containing no other ATG sequences) and the presumptive promoter area which displays a primary structure which is, for more than 110 bp, fully identical with the previously published promoter sequence of the so-called

'IFN-β_2' [19, 20]. This illustrates the identity of the two genes and demonstrates that we actually reached the physical 5' end of the chromosomal 26-kDa-protein gene.

The sequence information determined by combining cDNA information with genomic DNA information, is shown in Fig. 3. The total coding capacity now accounts for a polypeptide of 266 amino acids, which is far above the size estimate calculated for the 26-kDa cell-free translation product. However, it is quite possible that (a) small intron(s) in phase with the deduced reading frame is (are) still present in that part of the coding sequences which are derived from the genomic DNA.

As a matter of fact, on inspection of the coding area, it is remarkable that a region with a rather aberrant nucleotide composition emerges at the beginning of the gene. This region has a very high G + C content (70%) and an accumulation of the dinucleotide dC-dG which otherwise occurs rather infrequently in eukaryotic coding regions. In order to test the possible presence of a remaining intron in the genome-derived sequence part, a new reverse-transcription experiment was designed.

Reverse transcription and measurement of the intron size

For the purpose of extension with reverse transcriptase, total poly(A)-containing RNA, isolated from superinduced fibroblast cells, was hybridized to the complementary single-stranded DNA fragment of the *Hin*fI probe described previously (Fig. 1), localized in the front region of the cDNA clone. Primer extension should allow a more accurate deduction of the precise length of the cDNA transcripts synthesized and may avoid difficulties with internal stops along the entire length of the mRNA sequence.

An example of a primer-extension reaction is shown in Fig. 4. Starting with the *Hin*fI primer of 163 nucleotides, a number of prematurely terminated reverse-transcription products are found (one of which seems rather prominent) and only a fraction reaches full-length size. The length of this
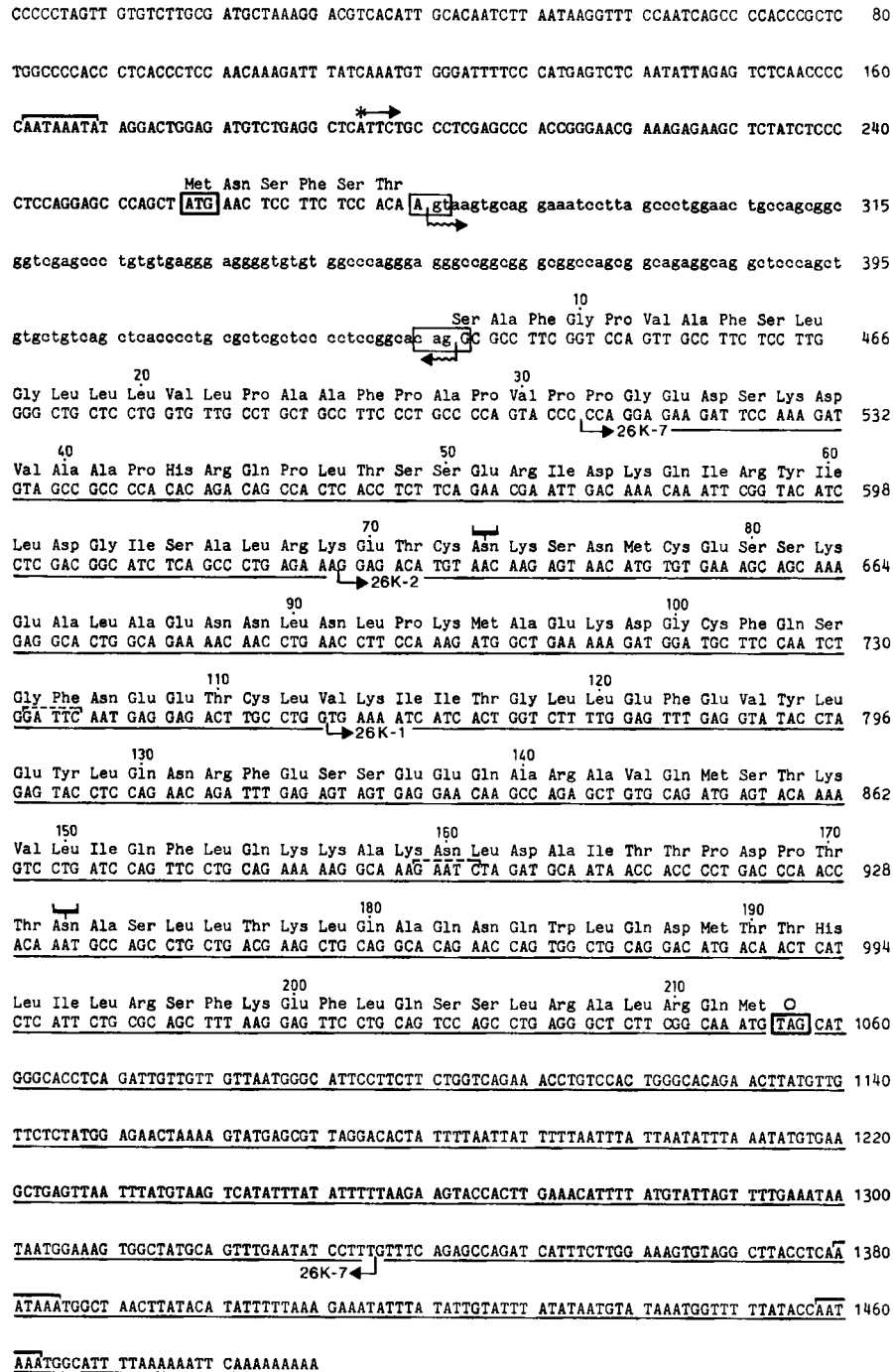
```
CCCCCTAGTT GTGTCTTGCG ATGCTAAAGG ACGTCACATT GCACAATCTT AATAAGGTTT CCAATCAGCC CCACCCGCTC  80

TGGCCCCACC CTCACCCTCC AACAAAGATT TATCAAATGT GGGATTTTCC CATGAGTCTC AATATTAGAG TCTCAACCCC  160

CAATAAATAT AGGACTGGAG ATGTCTGAGG CTCATTCTGC CCTCGAGCCC ACCGGGAACG AAAGAGAAGC TCTATCTCCC  240
                               * →

                          Met Asn Ser Phe Ser Thr
CTCCAGGAGC CCAGCT ATG AAC TCC TTC TCC ACA A gt aagtgcag gaaatcctta gccctggaac tgccagcggc  315

ggtcgagccc tgtgtgaggg aggggtgtgt ggcccaggga gggcggcgg gcggccagcg gcagaggcag gctcccagct  395

                                                        10
                                        Ser Ala Phe Gly Pro Val Ala Phe Ser Leu
gtgctgtcag ctcacccctg cgctcgctcc cctccggcac ag GC GCC TTC GGT CCA GTT GCC TTC TCC TTG  466

      20                                    30
Gly Leu Leu Leu Val Leu Pro Ala Ala Phe Pro Ala Pro Val Pro Pro Gly Glu Asp Ser Lys Asp
GGG CTG CTC CTG GTG TTG CCT GCT GCC TTC CCT GCC CCA GTA CCC CCA GGA GAA GAT TCC AAA GAT  532
                                                      └─►26K-7─────

      40                                    50                                      60
Val Ala Ala Pro His Arg Gln Pro Leu Thr Ser Ser Glu Arg Ile Asp Lys Gln Ile Arg Tyr Ile
GTA GCC GCC CCA CAC AGA CAG CCA CTC ACC TCT TCA GAA CGA ATT GAC AAA CAA ATT CGG TAC ATC  598

                              70                                        80
Leu Asp Gly Ile Ser Ala Leu Arg Lys Glu Thr Cys Asn Lys Ser Asn Met Cys Glu Ser Ser Lys
CTC GAC GGC ATC TCA GCC CTG AGA AAG GAG ACA TGT AAC AAG AGT AAC ATG TGT GAA AGC AGC AAA  664
                      └─►26K-2─────

                      90                                    100
Glu Ala Leu Ala Glu Asn Asn Leu Asn Leu Pro Lys Met Ala Glu Lys Asp Gly Cys Phe Gln Ser
GAG GCA CTG GCA GAA AAC AAC CTG AAC CTT CCA AAG ATG GCT GAA AAA GAT GGA TGC TTC CAA TCT  730

              110                                    120
Gly Phe Asn Glu Glu Thr Cys Leu Val Lys Ile Ile Thr Gly Leu Leu Glu Phe Glu Val Tyr Leu
GGA TTC AAT GAG GAG ACT TGC CTG GTG AAA ATC ATC ACT GGT CTT TTG GAG TTT GAG GTA TAC CTA  796
                              └─►26K-1─────

              130                                    140
Glu Tyr Leu Gln Asn Arg Phe Glu Ser Ser Glu Glu Gln Ala Arg Ala Val Gln Met Ser Thr Lys
GAG TAC CTC CAG AAC AGA TTT GAG AGT AGT GAG GAA CAA GCC AGA GCT GTG CAG ATG AGT ACA AAA  862

      150                                    160                                      170
Val Leu Ile Gln Phe Leu Gln Lys Lys Ala Lys Asn Leu Asp Ala Ile Thr Thr Pro Asp Pro Thr
GTC CTG ATC CAG TTC CTG CAG AAA AAG GCA AAG AAT CTA GAT GCA ATA ACC ACC CCT GAC CCA ACC  928

                                    180                                    190
Thr Asn Ala Ser Leu Leu Thr Lys Leu Gln Ala Gln Asn Gln Trp Leu Gln Asp Met Thr Thr His
ACA AAT GCC AGC CTG CTG ACG AAG CTG CAG GCA CAG AAC CAG TGG CTG CAG GAC ATG ACA ACT CAT  994

                      200                                    210
Leu Ile Leu Arg Ser Phe Lys Glu Phe Leu Gln Ser Ser Leu Arg Ala Leu Arg Gln Met  O
CTC ATT CTG CGC AGC TTT AAG GAG TTC CTG CAG TCC AGC CTG AGG GCT CTT CGG CAA ATG TAG CAT  1060

GGGCACCTCA GATTGTTGTT GTTAATGGGC ATTCCTTCTT CTGGTCAGAA ACCTGTCCAC TGGGCACAGA ACTTATGTTG  1140

TTCTCTATGG AGAACTAAAA GTATGAGCGT TAGGACACTA TTTTAATTAT TTTTAATTTA TTAATATTTA AATATGTGAA  1220

GCTGAGTTAA TTTATGTAAG TCATATTTAT ATTTTTAAGA AGTACCACTT GAAACATTTT ATGTATTAGT TTTGAAATAA  1300

TAATGGAAAG TGGCTATGCA GTTTGAATAT CCTTTGTTTC AGAGCCAGAT CATTTCTTGG AAAGTGTAGG CTTACCTCAA  1380
                               26K-7◄─┘

ATAAATGGCT AACTTATACA TATTTTTAAA GAAATATTTA TATTGTATTT ATATAATGTA TAAATGGTTT TTATACCAAT  1460

AAATGGCATT TTAAAAAATT CAAAAAAAAA
```

Fig. 3. *Complete sequence data of the 26-kDa-protein coding information and surronding areas and of the 26-kDa protein*. The entire nucleotide sequence is composed of cDNA information (underlined section) and of sequencing data derived from analysis of the 3800-bp *Eco*RI genomic DNA fragment (cf. Fig. 2). Arrows mark the 5' boundaries of the cDNA clones analyzed (cf. Fig. 1). Bars above the nucleotide sequence indicate regulatory signals, i.e. the TATA box and the two possible polyadenylation sites. The asterisk shows the presumptive 5' end of the 26-kDa-protein mRNA [19, 20]. The initiation and termination codons of the open reading frame are boxed. The small-letter sequence is used to indicate an intervening sequence; the wavy arrows and the boxes designate the corresponding donor and acceptor sites. The dashed brackets delineate the *Hin*fI cDNA fragment, of which the negative strand was used for the primer extension of the 26-kDa-protein mRNA as described in the text (cf. also Fig. 4). The primary structure of the 26-kDa polypeptide is derived from the coding information. The anchor symbol is used to designate the two potential Asn glycosylation sites

product amounts to about 520 nucleotides. In contrast, the theoretical length of the fully extended *Hin*fI fragment as deduced from the sequencing data (Fig. 3) would be around 700 nucleotides. We therefore concluded that our assumption regarding the presence of an additional intron in the genomic sequence preceding the cDNA information was correct and that the size of this intron is approximately 180 bp. We searched the established genome-derived 26-kDa-protein coding sequence for the presence of the prototypic nucleotide signal R'GT (a general formulation of a donor splice site as

deduced from a diversity of proposed consensus sequences [21 – 23]). The one closest to the putative mRNA start site follows the initiation codon ATG by 14 nucleotides. Looking further downstream for a possible acceptor splice site with general formula YAG'G [21 – 23], out of four such nucleotide sequences the one furthest downstream would, in combination with the previously mentioned donor site, give rise to an intron of 162 nucleotides. This intron size fits in well with the excess
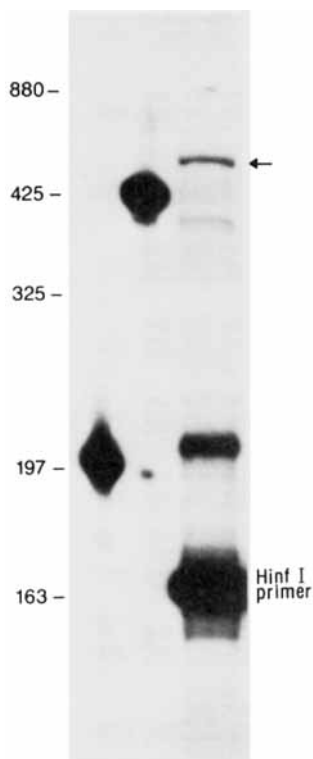
of sequence information as deduced from the reverse-transcription experiments. Furthermore, all other combinations of possible donor and acceptor splice signals present in this region of the DNA would result in introns which are much smaller, and, moreover, would affect the ongoing reading frame. In addition, the position of this 162-bp intron, as well as the surrounding exon sequences, exactly match the nucleotide-sequence data so far published for the so-called 'IFN-β_2' gene [20]. Furthermore, this intron almost completely eliminates the aberrant G + C-rich region mentioned above, and removes 10 pairs of the rarely used dinucleotide dC-dG from the coding region of the eukaryotic gene, thus restoring a more normal codon usage and nucleotide composition (the G + C content of the remaining 639-bp DNA coding region amounts to 50%).

As a result, the reading frame, after deletion of the intron (Fig. 3), has a total coding capacity for a 212-amino-acid polypeptide with an exact molecular mass of 23.7 kDa. This size correlates well with a 26-kDa protein product, as stated before.

DISCUSSION

The DNA sequence, as shown in Fig. 3, contains one unique reading frame, starting with an ATG codon at position 257 and terminating at position 1055; the coding region is followed by a rather long 3'-untranslated region of 424 nucleotides before it runs into the poly(A) tail. Remarkably, two so-called polyadenylation signals AATAAA are present in this region, separated by about 70 nucleotides, only the latter of which seems to be functional in determining poly(A) addition (although only a limited collection of cDNA clones were analyzed in this respect). A similar situation occurs in the gene of 'IFN-β_2' [24].

The region preceding the 26-kDa-protein gene is identical with the 'IFN-β_2' promoter previously described, the inducibility of which has been demonstrated in appropriate plasmid constructions [19]. Hence, there is little doubt that the sequence in front of the 26-kDa-protein gene has promoter activity. In this segment, the most remarkable sequence element is the TATA box, common to a large variety of eukaryotic polymerase II promoter areas. A tandem array of short, repetitive elements rich in purines is clearly absent from the 26-kDa-protein promoter area. As such a sequence arrangement (preceding the cap site by at most 120 bp) was
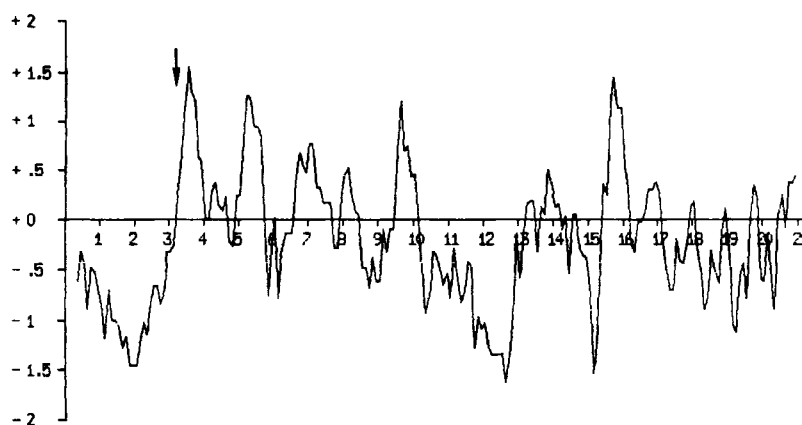

Fig. 4. *Gel fractionation of the primer extension reaction products on 26-kDa-protein mRNA*. The reverse-transcription reaction has been carried out using the single-stranded 163-bp *Hin*fI fragment as a primer; the extension products were fractionated on a 5% polyacrylamide gel containing 7 M urea. At both sides of the sample lane, markers were run to estimate the size of the extension products (the 325-bp and 880-bp references are not shown). The largest extension product, indicated by an arrow, corresponds to about 520 bp

Fig. 5. *Hydrophilicity plot of the 26-kDa polypeptide*. The hydrophilicity pattern of the 26-kDa polypeptide was calculated according to the principle developed by Hopp and Woods [28] and is plotted against the amino acid sequence. Note the hydrophobic 32-amino-acid N-terminal segment (the arrow marks the possible position of the signal cleavage site)

found to be responsible for inducibility by poly(I) · poly(C) in the corresponding promoters of IFN-β and IFN-α_1 [25 – 27], the lack of it from the 26-kDa-protein sequence agrees with the observation that cycloheximide treatment alone is sufficient for its induction [6].

The coding part of the 26-kDa-protein cDNA gene translates into a polypeptide of 212 amino acids, which is in the size range of a 26-kDa protein, as estimated for the *in vitro* translation product on a denaturing protein-sizing gel [5]. The base composition of the coding region has a fairly even distribution and the codon-usage table reveals no special preferences besides the fact that CG-containing codons are less frequently used than their synonymous homologues, and that the code words TTA for leucine and TAT for tyrosine are not used at all.

The 26-kDa protein itself is a pre-polypeptide, which means that a signal peptide of so far unknown length is cleaved off during *in vivo* maturation, giving rise to intermediates of lower molecular mass, as described previously [5]. Determination of the exact cleavage site will only be possible after amino acid analysis of the mature products. Nevertheless, a fairly hydrophobic, proline-rich peptide of about 32 amino acids (Fig. 5) is characteristic of the front sequence of the 26-kDa pre-polypeptide and may correspond to the presumptive signal peptide cleaved off during the secretion process.

Related proteins of higher molecular mass presumably correspond to additional modifications, such as glycosylation [5]. Two possible N-glycosylation sites may in fact be seen in the amino acid sequence (Fig. 3), but further studies at the protein level are needed to determine whether these sites are indeed substituted.

Previously, Revel [20] has pointed to homologous regions between IFN-β and the so-called 'IFN-β_2'. Analysis of the 26-kDa polypeptide indeed shows that seven out of thirteen amino acids in the area between codons 126 – 138 are identical with IFN-β codons in the region 28 – 43; somewhat further, a block of nine amino acids in the 26-kDa protein contains six identical amino acids within an IFN-β segment (codons 45 – 52). Whether these homologies bear any biological significance is not known, but this finding points once more to the identity between the 26-kDa protein and the so-called 'IFN-β_2'.

In a search for similarities between the 26-kDa polypeptide and any amino acid sequence so far published, we screened a protein database (containing 3309 individual peptide sequences) with the FASTP program of Lipman and Pearson [29]. However, none of the polypeptides scoring at Z values greater than 3 was an interferon species (Z values express the mathematical probability for the significance of a similarity). All homologies between these selected peptides and certain areas of the 26-kDa protein may be considered 'possibly significant' (Z > 3), but none of them reached the level of 'probably significant' (i.e. Z > 6), even after optimalization of their respective scores by the computer program. The same results were obtained whether dipeptide matches (ktup = 2) or individual entities (less stringent working mode, ktup = 1) were considered. This shows that the homology stretches between the interferon sequences present in the total data bank and the 26-kDa protein are of a lower significancy degree than the similarities between the 26-kDa protein and a variety of other, non-related peptides.

We then repeated the similarity search by the FASTP program to the IFN sequences only, present in the data bank (19 individual members, including human IFN-β) and now found the highest score of similarity with the IFN-β species

of mouse. To obtain an estimate of the statistical significance of the similarity, the query sequence (i.e. the 26-kDa peptide) was compared with randomly permuted versions of the potentially related amino acid sequence (mouse IFN-β) by an accompanying program RDF [29]. It was found that the similarity score for the authentic mouse IFN-β was lower than the values of randomly permuted versions and, furthermore, that these scores were below the limit of biological significance (i.e. Z > 3). Hence, we concluded that according to the FASTP program no convincing relationship could be detected between the 26-kDa protein and any interferon sequence available, as a variety of entirely different peptides from the database or randomly permuted versions of the most homologous IFN sequence score far better than any available IFN sequence as such.

Although the program used is very sensitive for detecting sequence similarities between different peptides, it can happen that non-contiguous regions of similarity are overlooked, due to the program selection for only the best similarity per entire peptide [30]. Therefore, we proceeded to another database search using the Los Alamos sequence analysis package [31], in which homology-symmetry searches are based on the powerful algorithm of Goad and Kanehisha [32]. According to this program, a negative weight value or 'distance' is attributed to each possible alignment of two sequences (the total value being the sum of all individual weights for matches, mismatches and deletions). The more negative the distance, the better the alignment and the more specific or significant it might be. When screening the protein sequence data bank PGtrans [33] for possible homologies to the 26-kDa peptide sequence, the best alignment with human IFN-β scored at a distance of − 53 and the alignment containing the homology region indicated by Revel [20] scored at −45; for comparison, in the same search the alignments of the 26-kDa-protein sequence with human IFN-αF and IFN-αH scored relatively better, i.e. at distances −65 and −61 respectively, while, in addition, the alignments between human IFN-β itself and different IFN-α species, for which a real structural homology of 29% at the amino acid level has already been described [34], scored at a distance of about − 300 or even more negative. As a matter of fact, in the homology search with the 26-kDa-protein sequence more than 5% of all the polypeptides present in the data bank scored between − 70 and − 60. From these results, we conclude that the homology between the 26-kDa peptide and any of the IFN sequences is below statistical significance.

Various lines of evidence, such as the mRNA induction conditions, protein size, immunological properties and, most unambiguously, nucleotide-sequence data, have proved beyond any doubt that the genes coding for the 26-kDa protein and for the so-called 'IFN-β_2', described by Revel and colleagues [2, 20], are identical. However, we were so far unable to ascribe unambiguously an antiviral activity to the 26-kDa-polypeptide-derived proteins isolated from the supernatant of induced cells or to the oocyte translation product [5]; also, we found that the induction requirements for the 26-kDa-protein gene are distinct from those necessary for the production of IFN-β [6]. Zilberstein et al. [35] did find biological IFN activity associated with the so-called 'IFN-β_2' and showed cross-neutralization of this β_2 activity using monospecific antiserum raised against pure IFN-β. It may be that their biological detection system was more sensitive. If, consequently, this molecule does indeed have a low specific antiviral activity it may be a representative of a new class of interferon rather than a novel fibroblast IFN subtype, as it

fails to show any significant structural homology with human fibroblast IFN-β. Nevertheless, the 26-kDa protein remains a remarkable cytokine because of its high induction levels obtained as a result of challenging fibroblast cells with IL1 or with cycloheximide. A major biological function has, however, still to be ascribed to this apparently important protein.

REFERENCES

1. Raj, N. B. & Pitha, P. (1980) *Proc. Natl Acad. Sci. USA 77*, 4918–4922.
2. Weissenbach, J., Chernajovsky, Y., Zeevi, M., Shulman, L., Soreq, H., Nir, U., Wallach, D., Perricaudet, M., Tiollais, P. & Revel, M. (1980) *Proc. Natl Acad. Sci. USA 77*, 7152–7156.
3. Sehgal, P. B. & Sagar, A. D. (1980) *Nature (Lond.) 288*, 95–97.
4. Sehgal, P. B. (1982) in *Interferon 4* (Gresser, I., ed.) pp. 1–22, Academic Press, New York.
5. Content, J., De Wit, L., Pierard, D., Derynck, R., De Clercq, E. & Fiers, W. (1982) *Proc. Natl Acad. Sci. USA 79*, 2768–2772.
6. Poupart, P., De Wit, L. & Content, J. (1984) *Eur. J. Biochem. 143*, 15–21.
7. Content, J., De Wit, L., Poupart, P., Opdenakker, G., Van Damme, J. & Billiau, A. (1985) *Eur. J. Biochem. 152*, 253–257.
8. Derynck, R., Content, J., DeClercq, E., Volckaert, G., Tavernier, J., Devos, R. & Fiers, W. (1980) *Nature (Lond.) 285*, 542–547.
9. Devos, R., van Emmelo, J., Contreras, R. & Fiers, W. (1979) *J. Mol. Biol. 128*, 595–619.
10. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) *Cell 15*, 1157–1174.
11. Benton, W. & Davies, R. (1977) *Science (Wash. DC) 196*, 180–182.
12. Rüther, U. (1982) *Nucleic Acids Res. 10*, 5765–5772.
13. Rigby, P. W. J., Dieckmann, M., Rhodes, C. & Berg, P. (1977) *J. Mol. Biol. 113*, 237–251.
14. Lillehaug, J. R., Kleppe, R. K. & Kleppe, K. (1976) *Biochemistry 15*, 1858–1865.
15. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol. 65*, 499–560.
16. Wu, R. (1970) *J. Mol. Biol. 51*, 501–521.
17. Volckaert, G., Tavernier, J., Derynck, R., Devos, R. & Fiers, W. (1982) *Gene 15*, 215–224.
18. Kozak, M. (1984) *Nucleic Acids Res. 12*, 857–872.
19. Chernajovsky, Y., Mory, Y., Chen, L., Marks, Z., Novick, D., Rubinstein, M. & Revel, M. (1984) *DNA 3*, 297–308.
20. Revel, M. (1983) in *Interferon 5* (Gresser, I., ed.) pp. 205–239, Academic Press, New York.
21. Breathnach, R., Benoist, C., O'Hare, K., Gannon, F. & Chambon, P. (1978) *Proc. Natl Acad. Sci. USA 75*, 4853–4857.
22. Seif, I., Khoury, G. & Dhar, R. (1979) *Nucleic Acids Res. 6*, 3387–3398.
23. Mount, S. M. (1982) *Nucleic Acids Res. 10*, 459–472.
24. Soreq, H., Sagar, A. D. & Sehgal, P. B. (1981) *Proc. Natl Acad. Sci. USA 78*, 1741–1745.
25. Fujita, T., Ohno, S., Yasumitsu, H. & Taniguchi, T. (1985) *Cell 41*, 489–496.
26. Ryals, J., Dierks, P., Ragg, H. & Weissmann, C. (1985) *Cell 41*, 497–507.
27. Goodbourn, S., Zinn, K. & Maniatis, T. (1985) *Cell 41*, 509–520.
28. Hopp, T. P. & Woods, K. R. (1981) *Proc. Natl Acad. Sci. USA 78*, 3824–3828.
29. Lipman, D. J. & Pearson, W. R. (1985) *Science (Wash. DC) 227*, 1435–1441.
30. George, D. G., Barker, W. C. & Hunt, L. T. (1986) *Nucleic Acids Res. 14*, 11–15.
31. Kaneshisa, M. I. (1982) *Nucleic Acids Res. 10*, 183–196.
32. Goad, W. B. & Kanehisa, M. I. (1982) *Nucleic Acids Res. 10*, 247–263.
33. Claverie, J.-M. & Sauvaget, I. (1985) *Nature (Lond.) 318*, 19.
34. Taniguchi, T., Mantei, N., Schwarzenstein, M., Nagata, S., Muramatsu, M. & Weissmann, C. (1980) *Nature (Lond.) 285*, 547–549.
35. Zilberstein, A., Ruggieri, R. & Revel, M. (1985) in *The interferon system 24* (Dianzani, F. & Rossi, G. B., eds) pp. 73–83, Raven Press, New York.

Note added in proof. We have recently shown that also tumor necrosis factor (TNF) can strongly induce 26-kDa-protein gene expression.