
The complete amino acid sequence of human fibroblast interferon as deduced using synthetic oligodeoxyribonucleotide primers of reverse transcriptase

M.Houghton, M.A.W.Eaton, A.G.Stewart, J.C.Smith, S.M.Doel, G.H.Catlin, H.M.Lewis, T.P.Patel, J.S.Emtage, N.H.Carey and A.G.Porter

Searle Research & Development, Lane End Road, High Wycombe, Bucks., UK

Received 30 May 1980

ABSTRACT

Using synthetic oligodeoxyribonucleotides to prime the transcription of interferon mRNA and cDNA, we recently determined the mRNA sequence coding for the 47 amino-terminal amino acids of mature human fibroblast interferon (1). From this sequence, we have now synthesised an oligodeoxyribonucleotide that is homologous with the mRNA sequence coding for amino acids 42-45 and used it as a primer to selectively transcribe an interferon cDNA template. The sequence of the newly synthesised DNA predicted the sequence of amino acids 48-109 in the interferon polypeptide. By repeating this process with one more primer, we have determined the complete amino acid sequence of mature human fibroblast interferon, a polypeptide of 166 amino acids.

INTRODUCTION

The interferons are cellular glycoproteins best known for their ability to induce in target cells a virus-resistant state (2), but are now also receiving attention because of their apparent anti-tumour activity (2,3). A necessary step in eventually understanding their mechanism of action is to determine their primary structure, an achievement that has been severely retarded by the tiny amounts of interferon produced in induced cells and by problems associated with their purification (4-7). Recently such problems have been partially overcome with the result that short N-terminal sequences for a few interferons have been determined (7-9). From these data, we deduced the structure of oligodeoxyribonucleotides that were capable of selectively priming the transcription of interferon mRNA and cDNA, and established the identity of the 47 N-terminal amino acids of mature human fibroblast interferon (F-IF) without having to resort to mRNA or protein purification methods (1).

Using similar priming methods, we now describe the elucidation of the complete amino acid sequence of this interferon, a polypeptide of 166 amino acids.

METHODS

Methods for the chemical synthesis of primers and their 5'-end labelling with (γ - ^{32}P) ATP, transcription of mRNA and cDNA, gel electrophoresis, nucleotide sequencing and induction of cells were all as described (1).

RESULTS

Based on our earlier sequence data for human fibroblast interferon mRNA (1), we synthesised the oligodeoxyribonucleotide primer 5' GAGGAGATTAAG 3' (P2) corresponding to amino acid positions 42-45 in the mature protein, and labelled the 5'-end to high specific activity with (γ - ^{32}P) ATP. It was annealed to single-stranded cDNA synthesised *in vitro* from total polyadenylated mRNA isolated from strain 17/1 human fibroblasts which had been induced to produce interferon under optimal conditions (1). Following incubation with reverse transcriptase and its substrates, the DNA products were electrophoresed through a 1.4% native agarose gel, and the gel was autoradiographed. A major band (cDNA_{P2}; Fig 1A) of about 850 base pairs was observed, eluted and the newly-synthesised DNA subjected to nucleotide sequence analysis. Part of the sequencing gel is shown in Fig. 2B from which it can be seen that the overlap with the interferon DNA sequence determined using the upstream primer 5' CTCTTTCCATG 3' (Fig. 2A; see also ref. 1) is 24 nucleotides. This proves that cDNA_{P2} is interferon cDNA. The one uncertain nucleotide in Fig. 2B occurs two nucleotides beyond the 3'-end of the primer (arrowed). This is assigned as an A residue based on the presence of an A in the corresponding position in the DNA synthesised with 5' CTCTTTCCATG 3' (arrowed in Fig. 2A). The observed heterogeneity in this position is more likely due to occasional loss of specificity of the chemical cleavages in the 5'-terminal 10-20 nucleotides (AP; unpublished observations) than to heterogeneity in the DNA.

A total of 193 nucleotides (excluding those of the primer) were identified from the sequencing gel of cDNA_{P2} from which the

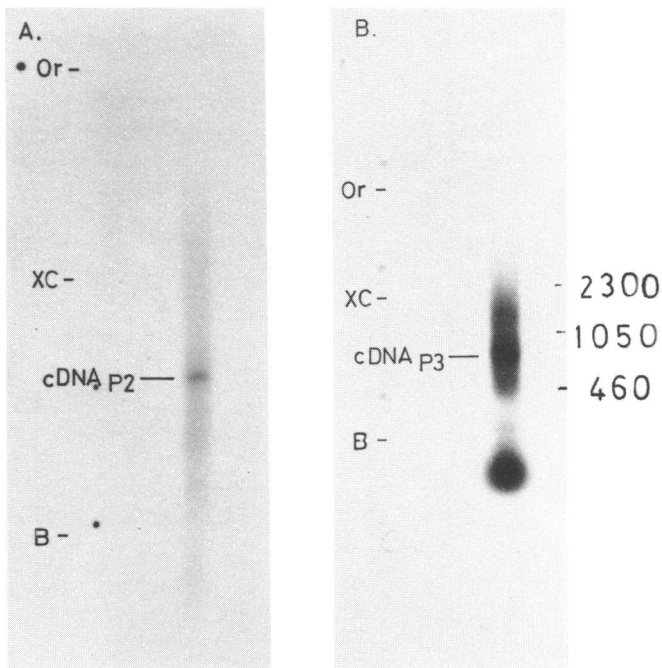


Figure 1 Agarose gel electrophoresis of primed DNA (1,17). A, cDNA_{P2} was made with (5'-³²P) GAGGAGATTAAG 3' (primer P2). B, cDNA_{P3} was made with (5'-³²P) CCTGGAAGAAA 3' (primer P3). The gels were run at 40V for 16-18 hr and autoradiographed at 4°C. XC and B refer to xylene cyanol and bromophenol blue dye markers, and in B the positions of known DNA markers are given in base-pairs.

sequence of a further 62 amino acids of the protein was deduced (positions 48-109; Fig. 3).

Human fibroblast interferon is a glycoprotein of $M_r \sim 19,000-20,000$ (4-6, 14) with a carbohydrate content of at least 20% (14). From these data, it was calculated that the mature protein contains about 135 amino acids, and hence we required to establish the identity of about 26 more amino acids (78 nucleotides). To achieve this, we next made the oligodeoxyribonucleotide 5' CCTGGAAGAAA 3' (P3), which contains the codons for amino acid positions 102-104 (Fig. 3). As before, the new primer was labelled and used to prime transcription on an interferon cDNA template. Once again, agarose gel electro-

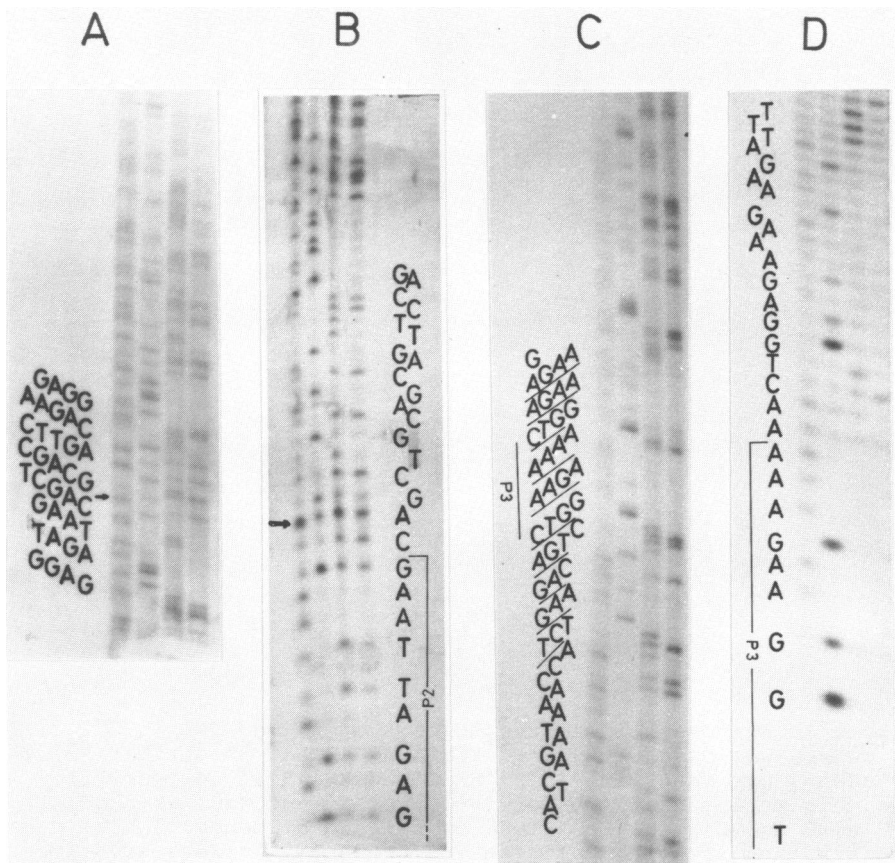


Figure 2 Autoradiographs of sequencing gels. The cleavage specificities are from left to right; AC, G, TC, C. The F-IF DNA bands (cDNA_{P2} and cDNA_{P3}; Fig. 1) were eluted from the agarose gels (1), subjected to the partial chemical degradations (18) and the products were resolved in 8% polyacrylamide sequencing gels (12). A. Sequence deduced using 5' CTCTTCCATG 3' as primer (1) on which synthesis of 5' GAGGAGATTAAG 3' (P₂) was based. The clear A residue which is uncertain in B is arrowed. B. 5'-terminal sequence of cDNA_{P2} showing part of the sequence deduced previously using the primer 5' CTCTTCCATG 3' (see ref. 1. and Fig. 3). C. Sequence in cDNA_{P2} on which the synthesis of the primer 5' CCTGGAAGAAA 3' (P₃) was based. D. 5'-terminal sequence of cDNA_{P3}.

phoresis revealed a major band of about 850 base pairs (cDNA_{P3}; Fig. 1B), although this product was not as pure as transcripts derived from other primers owing to a heavy background smear.

Nonetheless, when eluted this DNA band gave a clear nucleotide sequence which showed that the 15 nucleotides adjacent to the primer (Fig. 2D) were identical to the corresponding 15 nucleotides deduced using the previous primer 5' GAGGAGATTAAG 3' (Fig. 2C). The rest of the unambiguous sequence derived from cDNA_{P3} extends for eight nucleotides beyond an in-phase termination codon (UGA) in the mRNA, establishing the sequence of the C-terminal 57 amino acids of human F-IF, making a total of 166 amino acids altogether (Fig. 3). The termination codon occurs at nucleotides 633-635 relative to the presumed 5'-end of the mRNA. Assuming the mRNA is about 850 nucleotides in length, it may be calculated that there are an additional ~200-215 nucleotides at the 3'-end which are presumably non-coding.

With all three primers (P1-P3; Fig. 3) the size of the interferon gene product always corresponded to the expected size of full-length double-stranded interferon genes (about 850 base pairs; (10)), when electrophoresed through native agarose gels. On the other hand, when electrophoresed through denaturing polyacrylamide gels, the size of the labelled product correlated with the position in the cDNA template where the primer annealed (MH-unpublished). These effects can be explained by assuming that as well as acting as template for each oligodeoxyribonucleotide primer, the single stranded interferon cDNA self-primed by virtue of the 3'-end looping back on itself (11,12). The transcript is then elongated up to the position of the oligodeoxyribonucleotide-primed transcript where it is halted, thus giving rise to a full-length double-stranded interferon "gene" in which the 3'-end of the self-primed transcript and the 5'-end of the oligodeoxyribonucleotide-primed transcript are not covalently joined.

DISCUSSION

Clearly, the methods used in this and our previous report (1) provide a powerful way of sequencing rare mRNA species that are present in a total mRNA population, without having to resort to either nucleic acid purification or cloning methods.

We have shown that the DNA sequence determined using each primer is interferon-specific as there is sufficient overlap between sequences deduced with adjacent primers (Fig. 3). Also,

Nucleic Acids Research

Next to 5'-cap? 20 40

Coding strand 5'-UUCUAACUGCAACCUUUCGAAGCCUUUGCUCUGGCACAACAGGUAGUAGGCCGACACUGU
 5'-untranslated

60 80 100 120
 UCGUGUUGUCAAC AUG ACC AAC AAG UGU CUC CUC CAA AUU GCU CUC CUG UUG UGC UUC UCC
 Met-Thr-Asn-Lys-Cys-Leu-Leu-Gln-Ile-Ala-Leu-Leu-Leu-Cys-Phe-Ser-
Precursor peptide -15 -10

P1 START 140 160
 ACU ACA GCU CUU UCC AUG AGC UAC AAC UUG CUU GGA UUC CUA CAA AGA AGC AGC AAU
 Thr-Thr-Ala-Leu-Ser-Met-Ser-Tyr-Asn-Leu-Leu-Gly-Phe-Leu-Gln-Arg-Ser-Ser-Asn-
 -5 -1 N-terminus mature F-IF 10

180 200 220
 UUU CAG UGU CAG AAG CUC CUG UGG CAA UUG AAU GGG AGG CUU GAA UAC UGC CUC AAG
 Phe-Gln-Cys-Gln-Lys-Leu-Leu-Trp-Gln-Leu-Asn-Gly-Arg-Leu-Glu-Tyr-Cys-Leu-Lys-
20 30

P2 START
240 260 280
 GAC AGG AUG AAC UUU GAC AUC CCU GAG GAG AUU AAG CAG CUG CAG CAG UUC CAG AAG
 Asp-Arg-Met-Asn-Phe-Asp-Ile-Pro-Glu-Glu-Ile-Lys-Gln-Leu-Gln-Gln-Phe-Gln-Lys-
40 50

P1 END
300 320 340
 GAG GAC GCC GCA UUG ACC AUC UAU GAG AUG CUC CAG AAC AUC UUU GCU AUU UUC AGA
 Glu-Asp-Ala-Ala-Leu-Thr-Ile-Tyr-Glu-Met-Leu-Gln-Asn-Ile-Phe-Ala-Ile-Phe-Arg-
60 70

360 380 400
 CAA GAU UCA UCU AGC ACU GGC UGG AAU GAG ACU AUU GUU GAG AAC CUC CUG GCU AAU
 Gln-Asp-Ser-Ser-Ser-Thr-Gly-Trp-Asn-Glu-Thr-Ile-Val-Glu-Asn-Leu-Leu-Ala-Asn-
80 90

P3 START
420 440 460
 GUC UAU CAU CAG AUA AAC CAU CUG AAG ACA GUC CUG GAA GAA AAA CUG GAG AAA GAA
 Val-Tyr-His-Gln-Ile-Asn-His-Leu-Lys-Thr-Val-Leu-Glu-Glu-Lys-Leu-Glu-Lys-Glu-
100

P2 END
480 500
 CAU UUC ACC AGG GGA AAA CUC AUG AGC AGU CUG CAC CUG AAA AGA UAU UAU GGG AGG
 Asp-Phe-Thr-Arg-Gly-Lys-Leu-Met-Ser-Ser-Leu-His-Leu-Lys-Arg-Tyr-Tyr-Gly-Arg-
 110 120

520 540 560
 AUU CUG CAU UAC CUG AAG GCC AAG GAG UAC AGU CAC UGU GCC UGG ACC AUA GUC AGA
 Ile-Leu-His-Tyr-Leu-Lys-Ala-Lys-Glu-Tyr-Ser-His-Cys-Ala-Trp-Thr-Ile-Val-Arg-
130 140

continued...

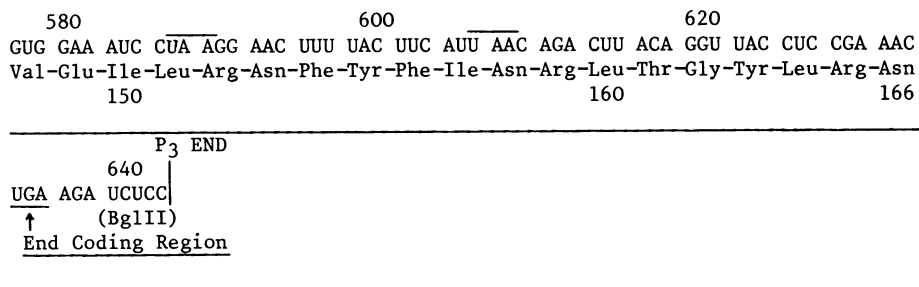


Figure 3 Nucleotide sequence of F-IF mRNA and predicted amino acid sequence of human F-IF. Nucleotides are numbered from the presumed 5'-terminal "cap" of the mRNA, while amino acids are numbered from the N-terminus of the mature protein. Out-of-phase nonsense codons are indicated by a line above the sequence. The nucleotides corresponding to the 5'-end of each primer are marked P₁-P₃ START and the distances sequenced with each primer are indicated by P₁-P₃ END. (P₁ corresponds to IF II in ref. 1).

as previously mentioned, in native agarose gels the size of the double-stranded interferon DNA was always the same. Although the coding region was sequenced only on one strand, we are confident that the number of nucleotides is correct as the sequencing gel patterns were clear with no visible "compression" of bands (13). The presence of stop codons in two of the three reading frames throughout the coding region supports this assertion.

The amino acid composition of mature human F-IF is shown in Table 1 from which the molecular weight of the unglycosylated polypeptide can be estimated to be 20,014 Daltons. The values of 19,000-20,000 Daltons for purified human F-IF (4-6, 14) therefore appear to be slight underestimates as they include a contribution equivalent to 4,000 Daltons from the carbohydrate moiety (14). The location of the carbohydrate side chain(s) is not known, although the amino acid sequence Asn-Glu-Thr (positions 80-82; Fig. 3) is a potential recognition site for attachment of an oligosaccharide via the amino group on the asparagine (15). Almost half of the interferon polypeptide consists of amino acids with hydrophobic side chains (particularly phenylalanine, tyrosine, leucine and isoleucine) while glycine

Table 1

Amino acid composition in residues per mole of mature human 17/1 F-IF deduced from the amino acid sequence in Fig. 3 with the published values determined by amino acid analysis of purified human C-10 (6) and FS-4 (7) F-IF in brackets.

	C-10	FS-4		C-10	FS-4
Asp 5	(20.0)	(18.9)	Val 5	(7.6)	(6.0)
Asn 12			Met 4	(0.5)	(2.9)
Thr 7	(8.0)	(6.8)	Ile 11	(9.8)	(9.0)
Ser 9	(11.5)	(10.5)	Leu 24	(26.2)	(20.4)
Glu 13	(26.9)	(27.0)	Tyr 10	(3.2)	(7.5)
Gln 11			Phe 9	(7.6)	(9.4)
Pro 1	(4.3)	(2.7)	His 5	(4.5)	(4.9)
Gly 6	(5.3)	(7.8)	Lys 11	(11.9)	(11.6)
Ala 6	(9.1)	(10.0)	Arg 11	(8.5)	(10.9)
Cys 3	(8.0)	(1.7)	Trp 3	(0)	(1.0)

is unusually low (six residues).

Our sequence of strain 17/1 fibroblast interferon is identical to that of the fibroblast cell line DIP 2 determined from cloned DNA (Taniguchi *et al.*, Gene, in press). In the corresponding nucleotide sequence the triplet coding for tyrosine at amino acid position 30 (Fig. 3) is UAC in strain 17/1 interferon mRNA and UAU in the case of DIP 2. This difference could be due to a wrong nucleotide being inserted during transcription or cloning, or more likely, it could reflect a silent mutation. Apart from this difference, the nucleotide sequences are identical.

The nucleotide sequence in Fig. 3 was derived from transcripts of non-purified polyadenylated interferon mRNA and there was no clear evidence of heterogeneity in the sequencing gels. Also, as mentioned above, the double-stranded interferon genes were similar in size in native agarose gels whichever primer was used. These observations indicate that we have detected one type of interferon mRNA with the primers. However, we cannot rule out either that some of the fainter bands in the agarose gels (e.g. Fig. 1B) contain cDNA derived from minor interferon mRNA species or the existence of other interferon cDNA species (16) that fail to bind the primers.

Human fibroblast interferon (see also Taniguchi *et al.*,

Gene, in press) and leukocyte interferon (Mantel et al., Gene, in press) are the first interferons to be completely sequenced. It will now be possible to look for the presence of conserved regions that may indicate important common functions (Taniguchi et al., in press).

ACKNOWLEDGEMENTS

We wish to thank Dr. J. Pardon for considerable help with computer programming, Drs. J. Birch and T. Cartwright for supplying 17/1 fibroblast cells and Dr. A. J. Hale for the provision of excellent research facilities.

REFERENCES

1. Houghton, M., Stewart, A.G., Doel, S.M., Emtage, J.S., Eaton, M.A.W., Smith, J.C., Patel, T.P., Lewis, H.M., Porter, A.G., Birch, J.R., Cartwright, T. and Carey, N.H. (1980) *Nucleic Acids Res.* 8, 1913-1931.
2. Stewart, W.E.II and Lin, L.S. (1979) *Pharmac. Ther.* 6, 443-512.
3. Stewart, W.E.II (1979) *The Interferon System* (Springer, Berlin).
4. Knight, E. (1976) *Proc. Natl. Acad. Sci. USA* 73, 520-523.
5. Berthold, W., Tan, C. and Tan, Y.H. (1978) *J. Biol. Chem.* 253, 5206-5212.
6. Tan, Y.H., Barakat, F., Berthold, W., Smith-Johannsen, H. and Tan, C. (1979) *J. Biol. Chem.* 254, 8067-8073.
7. Knight, E., Hunkapillar, M.W., Korant, B.D. and Hardy, R.W.F. (1980) *Science* 207, 525-526.
8. Zoon, K.C., Smith, M.E., Bridgen, P.J., Anfinsen, C.B., Hunkapillar, M.W. and Hood, L.E. (1980) *Science* 207, 527-528.
9. Taira, H., Broeze, R.J., Jayaram, B.M., Lengyel, P., Hunkapillar, M.W. and Hood, L.E. (1980) *Science* 207, 528-529.
10. Sehgal, P.B., Lyles, D.S. and Tamm, I. (1978) *Virology* 89, 186-198.
11. Seeburg, P.H., Shine, J., Martial, J.A., Baxter, J.D. and Goodman, H.M. (1977) *Nature* 270, 486-494.
12. Porter, A.G., Barber, C., Carey, N.H., Hallewell, R.A., Threlfall, G. and Emtage, J.S. (1979) *Nature* 282, 471-477.
13. Sanger, F., Nicklen, S. and Coulson, A.R. (1977) *Proc. Natl. Acad. Sci. USA* 74, 5463-5467.
14. Havell, E.A., Yamazaki, S. and Vilček, J. (1977) *J. Biol. Chem.* 252, 4425-4427.
15. Neuberger, A., Gottschalk, A., Marshall, R.D. and Spiro, R. G. (1972) *The glycoproteins; their composition, structure and function*, 2nd ed., pt A, pp 450-490 (Elsevier, Amsterdam).
16. Sehgal, P.B., and Sagar, A. (March 1980) (personal communication).

17. Sharp, P.A., Sugden, B. and Sambrook, J. (1973) *Biochemistry* 12, 3055-3063.
18. Maxam, A. and Gilbert, W. (1977) *Proc. Natl. Acad. Sci USA* 74, 560-564.