Proc. Natl. Acad. Sci. USA
Vol. 81, pp. 7787–7791, December 1984
Biochemistry

Sequence of the initiation factor IF2 gene: Unusual protein features and homologies with elongation factors

(DNA sequence/sequence homology/protein secondary structure)

CHRISTINE SACERDOT*, PHILIPPE DESSEN†, JOHN W. B. HERSHEY‡, JACQUELINE A. PLUMBRIDGE*, AND MARIANNE GRUNBERG-MANAGO*

*Institut de Biologie Physico-Chimique, 13 rue Pierre et Marie Curie, 75005 Paris, France; †Laboratoire de Biochimie, Ecole Polytechnique, 91128 Palaiseau, France; and ‡Department of Biological Chemistry, School of Medicine, University of California, Davis, CA 95616

Contributed by Marianne Grunberg-Manago, August 29, 1984

ABSTRACT **The gene for protein synthesis initiation factor IF2 in *Escherichia coli*, *infB*, is located downstream from *nusA* on the same operon. We sequenced about 3 kilobases of DNA beginning within *nusA* and including the entire *infB* structural gene plus another 392 bases downstream. This region contains no obvious strong promoter signals, but a possible transcriptional termination or pausing site occurs downstream from *infB*. The putative initiator codon for IF2α (97,300 daltons) is AUG; that for IF2β (79,700 daltons) is GUG, located 471 bases downstream in the same reading frame. The codon usage for IF2 is typical of other highly expressed proteins in *E. coli* and suggests that IF2 mRNA is efficiently translated. IF2α contains two adjacent regions (residues 104–155 and 167–214) that are rich in alanine and charged amino acids and that show striking periodicities in their sequences. These regions may alternate between flexible and helical conformations, thereby drawing together the NH_2-terminal and COOH-terminal globular domains of the factor as IF2 interacts with ribosomes or tRNA. Certain regions of the DNA and protein sequences of IF2 share strong homologies with elongation factor EF-Tu and lesser homology with EF-G. In particular, a region of EF-Tu implicated in GTP binding contains sequences and secondary structure that are conserved in IF2. The homologies indicate that the genes for IF2 and the elongation factors are derived at least in part from a common ancestor.**

Initiation of protein synthesis in *Escherichia coli* is promoted by three initiation factors called IF1, IF2, and IF3. IF2 is involved in at least two steps in the initiation pathway: the binding of formylmethionyl-tRNA (fMet-tRNA) to 30S ribosomal subunits and the hydrolysis of GTP when the 50S subunit joins the 30S initiation complex to form the 70S complex (for reviews, see refs. 1 and 2). The role of IF2 in initiation may be compared to that of elongation factor EF-Tu, which promotes the binding of aminoacyl-tRNAs to 70S ribosomes and hydrolyzes GTP during elongation. Whereas EF-Tu forms a stable ternary complex with aminoacyl-tRNA and GTP in the absence of ribosomes, a similar complex containing IF2, fMet-tRNA, and GTP has not been isolated. However, IF2 specifically protects fMet-tRNA from spontaneous hydrolysis of the aminoacyl ester bond (3).

To obtain more detailed information on the function of IF2, we isolated the *infB* gene for IF2 (4). The gene maps at 68.5 min on the *E. coli* map (4) and is cotranscribed with the $tRNA_{f2}^{Met}$ gene and with *nusA*, whose product is involved in transcription termination (5). *InfB* codes for two molecular mass forms of IF2, called IF2α (97.3 kDa) and IF2β (79.7 kDa), both of which are found in *E. coli* cells (6). Here we report the DNA sequence of the *infB* gene and its proximal surrounding regions and describe interesting features of the primary structure of IF2.

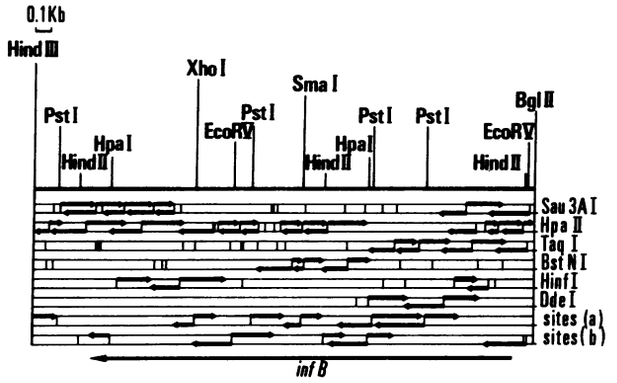

FIG. 1. Fine restriction map of the *Hin*dIII–*Bgl* II DNA fragment containing the *infB* gene. *E. coli* strain CSR603 was used to host plasmid pB16-1 (7). Plasmid DNA was isolated from cleared cell lysates according to Clewell (8). The horizontal arrows starting from the cleavage sites show the nucleotides sequenced by the Maxam and Gilbert method (9) on each DNA strand as described (10). Sites (a) are *Hin*dIII, *Xho* I, *Sma* I, and *Pst* I sites. Sites (b) are *Hpa* I, *Hin*dIII, and *Eco*RV sites. The arrow labeled *infB* at the bottom corresponds to the region coding for IF2α.

RESULTS

DNA Sequence of the *infB* Gene. A plasmid pB16-1 was constructed that contains a 3.2-kilobase (kb) *Hin*dIII–*Bgl* II fragment cloned into the *Hin*dIII and *Bam*HI sites of the tetracycline resistance gene of pBR322 and that expresses both IF2α and IF2β (7). The insert was sequenced by the Maxam and Gilbert method by using the strategy shown in Fig. 1, and the complete sequence is shown in Fig. 2. The sequence contains a portion of the *nusA* gene, the *nusA–infB* intercistronic region, the whole *infB* structural gene, and a 392-base sequence downstream from the gene. Several interesting features of this sequence are discernible. The *infB* structural gene starts 21 base pairs (bp) downstream from the second UAA stop codon of *nusA* and therefore is in the same reading frame. The putative AUG initiator codon (positions 142–144) is preceded by an A-A-G-G-A Shine–Dalgarno (11) sequence (132–136). Following this AUG is an open reading frame of 2667 bp that codes for a protein of 890 amino acids. The calculated mass of the protein is 97.3 kDa, which is compatible with reported molecular masses for IF2α (1). A GUG initiator codon (613–615) for IF2β was identified that lies 471 bp downstream from the IF2α initiator codon and generates a protein of 79.7 kDa in the same reading frame as IF2α. This codon is preceded by a GGA (600–602) and an AAG (605–

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: IF, initiation factor(s); EF, elongation factor(s); kb, kilobase(s); bp, base pair(s).

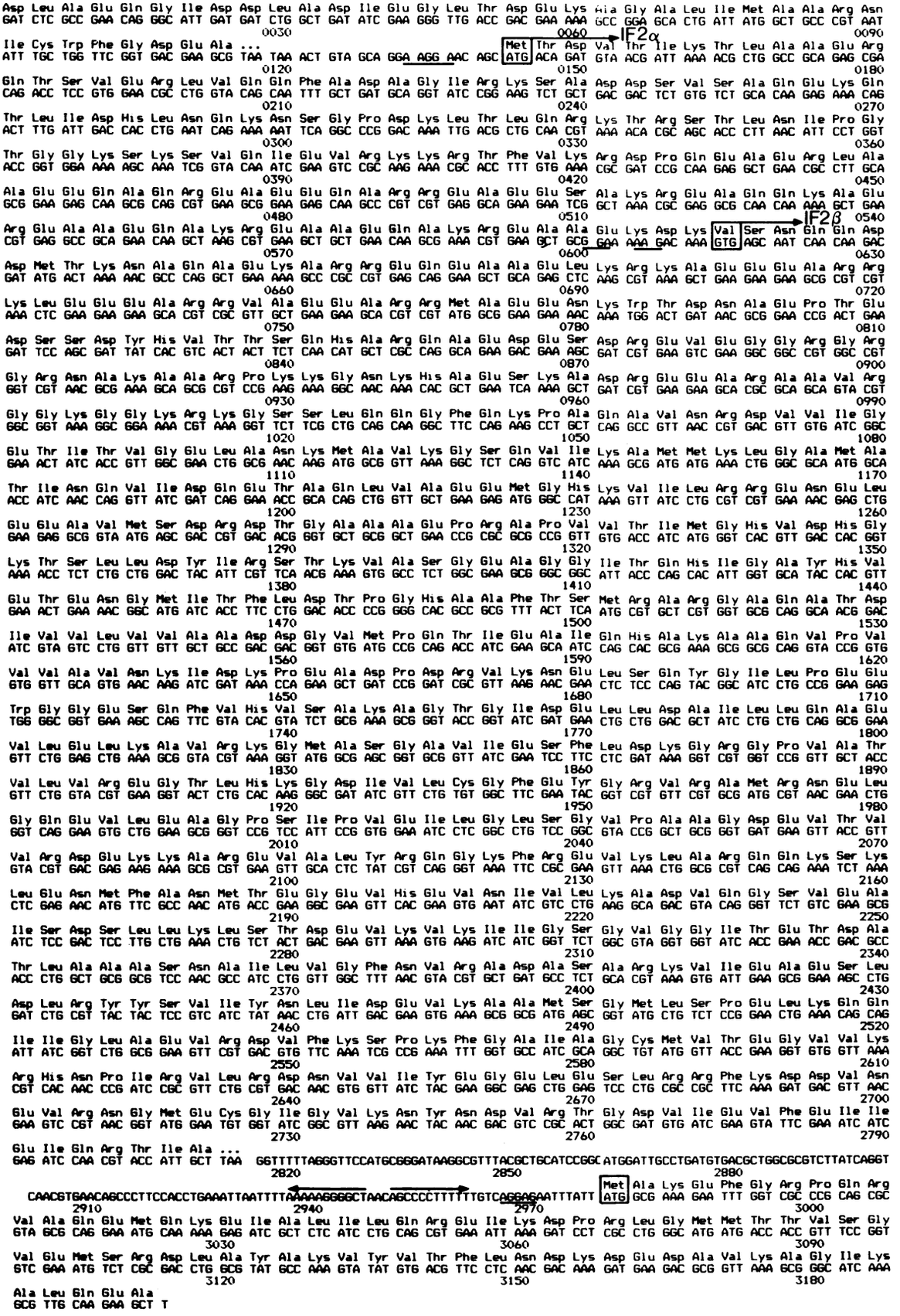

FIG. 2. Nucleotide sequence of the *infB* gene. The DNA sequence of the *Bgl* II–*Hin*dIII fragment includes the end of *nusA*, all of *infB*, the beginning of another open reading frame, and intercistronic regions. Corresponding amino acid sequences are shown above the DNA anti-sense strand. Putative Shine–Dalgarno sequences are underlined before each initiator codon (in boxes) for IF2α, IF2β, and the last open reading frame. An extensive region of dyad symmetry corresponding to a possible transcription terminator is indicated by divergent horizontal arrows above the DNA sequence.

607) sequence, one of which could be a weak Shine–Dalgarno sequence. The two initiation sites for IF2 were deduced from the NH_2-terminal amino acid sequences of IF2α and IF2β, from dipeptide initiation assays (12) with *infB*-containing plasmids, and from analyses of fused proteins (to be reported elsewhere).

Codon usage in *infB* resembles that in *E. coli* proteins that are highly expressed (13). Codon usage can be compared by calculating for each amino acid the ratio of the use frequency of a codon relative to the frequency of the most used codon in a group of highly expressed genes (Table 2 in ref. 13). Such ratios are averaged to give e, the coefficient of expressivity. This coefficient gives a measure of the expressivity of a coding frame independent of its amino acid composition. It shows that *infB* is quite strongly expressed ($e = 0.803$), approaching the values for the highly expressed genes *tufA* ($e = 0.904$), *tufB* ($e = 0.892$), and *fus* ($e = 0.877$) and the genes of some ribosomal proteins ($0.597 < e < 0.900$). The part of *infB* that corresponds to IF2β has a slightly higher expressivity ($e = 0.831$) than the whole structural gene, whereas the coding region extending from the first initiator codon (AUG, positions 142–144) to the second one (GUG, 613–615) is significantly less strongly expressed ($e = 0.673$), containing more codons of rare isoacceptor tRNA species. This coding region may be less efficiently translated and thus promote initiation at the IF2β initiator codon.

Downstream from the termination codon in *infB* (UAA, positions 2812–2814), we find a sequence (2936–2960) that can form a stable hairpin and loop structure ($\Delta G = -20$ kcal; 1 cal = 4.184 J). Moreover, this hairpin structure is followed by a T-rich region (T-T-T-T-T-T-G-T, 2956–2963) and thus resembles documented transcription terminators (14). The efficiency of this transcriptional termination or pausing site remains to be investigated. The structure is followed immediately by an open reading frame starting at an AUG (2978–2980) codon and preceded by a good Shine–Dalgarno sequence (A-G-G-A-G, 2965–2969). This open reading frame, which codes for a protein of at least 8.5 kDa from the sequence reported here, may correspond to the beginning of the 15-kDa protein described by Kurihara and Nakamura (15).

The Protein Sequence of IF2. The overall amino acid composition of the IF2α protein predicted from the DNA sequence shows few differences compared to the average composition of *E. coli* proteins (16): there are fewer cysteine residues (0.33/1.1%), aromatic residues (phenylalanine, 1.7/3.6%; tryptophan, 0.2/1.2%; tyrosine, 1.2/2.6%), and twice as many glutamic acid residues (11.1/5.5%). We also analyzed the amino acid sequence for internal homologies by using the method of Staden (17). As shown in Fig. 3*A*, two adjacent regions have highly unusual primary structures.

The first region (residues 104–155; nucleotides 451–606) has a very abnormal amino acid composition, made up of essentially only five different amino acids: 16 alanine, 16 glutamic acid, 5 lysine, 7 arginine, 7 glutamine (and 1 serine). The strict repetition of alanine residues every 4 amino acids, the pattern Arg-Glu-Ala repeated six times, and a periodicity of 8 residues are quite remarkable. A helical secondary structure is predicted for this region by the empirical method of Garnier *et al.* (19). The placement of alanine every 4 residues suggests that a 4-fold helix (e.g., ω-helix) is a possibility. The juxtaposition of amino acid residues in both α-helical and ω-helical conformations is shown in Fig. 4. The regularity of the alanine residues and the distribution of charged groups to maximize neutralization are particularly striking.

The second region concerns residues 167–214 (nucleotides 640–783). Its amino acid composition is also rich in alanine, glutamic acid, lysine, and arginine residues. The pattern Glu-Glu-Ala-Arg-Arg is strictly repeated three times, and a periodicity of 8 residues is seen (Fig. 3*B*). A helical structure is predicted (19), and α- and ω-helices are shown in Fig. 4. The pattern of alternating groups of positive and negative charges is most apparent in the ω-helix conformation. In addition, the first and second regions share amino acid (50%) and nucleotide (57%) sequence homologies. The functional implications of these unusual sequences are discussed below.

Homologies in Primary and Secondary Structure Between IF2 and Elongation Factors EF-Tu and EF-G. Because of similarities in the functional roles of IF2 and EF-Tu, we compared their DNA and amino acid sequences. Two regions of *tufA* (20) in the proximal and middle portions of the gene show extensive homologies with middle portions of *infB* (bases 1318–1384 and 1513–1644). The DNA sequences are aligned in Fig. 5 and the corresponding amino acids are shown along with those for EF-G (21). The first region concerns residues Val-393 to Ile-411 of IF2 and residues Val-14 to Ile-31 near the NH_2-terminus of EF-Tu. Of the 19 IF2 residues, 13 correspond exactly to those in EF-Tu, and a string of 8 residues is identical in both proteins. Even more striking is the conservation of the DNA sequence in this subregion, where 90% homology is observed in a 30-nucleotide sequence.

The second region concerns residues Gly-458 to Asp-501 of IF2 and residues Gly-94 to Asp-138 of EF-Tu. Striking homologies exist from Asp-463 to Glu-481 of IF2: this subregion exhibits 68% homology at the level of both amino acids and nucleotides. There are also minor homologies of IF2 and EF-Tu with EF-G; 6 residues and 3 dipeptides are conserved

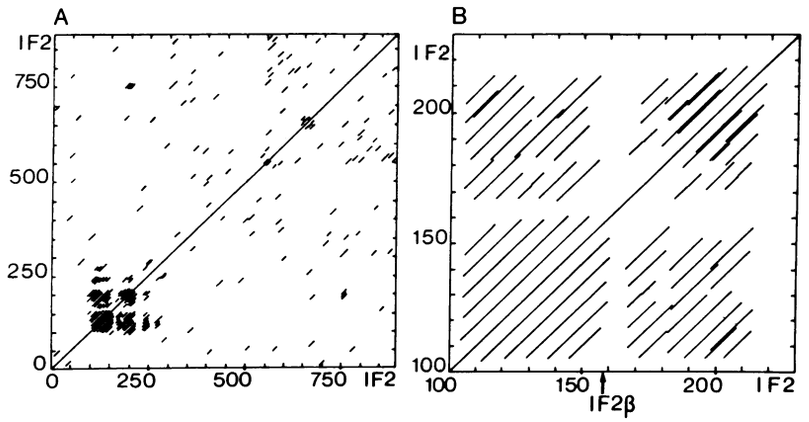

FIG. 3. Internal homologies of IF2α. (*A*) Comparison matrix of internal repetitions of the IF2 amino acid sequence. A proportional matching method based on that of Staden (17) was used with a span of 11 residues, a score at the 1% level of expectation, and the score matrix MDM78 of Dayhoff *et al.* (18). (*B*) An expanded region of *A*.

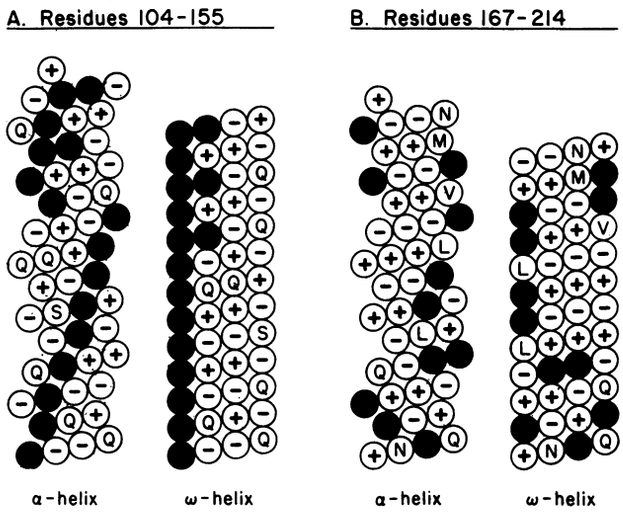

FIG. 4. Highly repetitious IF2 amino acid sequences. Shown is the amino acid sequence of IF2α from residues 104 to 155 and from residues 167 to 214 as open representations of α- and ω-helical conformations. Alanine residues are solid circles; + represents arginine and lysine residues; − represents glutamic acid and aspartic acid residues; Q is glutamine, S is serine, N is asparagine, L is leucine, V is valine, and M is methionine.

in all three factors. The secondary structure of this region of the EF-Tu protein has been reported recently (22): an α-helical region is flanked by two β-strands as shown in Fig 5. The junction between β-strand 3 and α-helix II (EF-Tu residues 110–113) is involved in the binding of GTP (22), and 3 of 4 residues are identical in IF2. The junction between α-helix II and β-strand 4 consists of the dipeptide Val-Pro (EF-Tu residues 127–128), which is conserved in all three factors. Moreover, the secondary structure predicted (19) for IF2 in this part of the protein also is similar to the known EF-Tu structure. The region from Ile-464 to Ala-471 consists exclusively of hydrophobic residues and is predicted to occur in a β-strand structure, similar to β-strand 3 of EF-Tu. The IF2 regions homologous to the α-helix II and β-strand 4 of EF-Tu are predicted to occur as an α-helix followed by a β-strand.

DISCUSSION

We have determined the DNA sequence of the *infB* gene, which provides both the primary structure of IF2 and the structural basis for evaluating the expression of the gene. Our sequence extends the DNA sequence of the $tRNA_{f2}^{Met}$–*nusA–infB* operon (5) by about 3 kb. No strong consensus promoter-like sequence has been found in our sequence, but a putative termination or pause site is found distal to the *infB* gene. Further work is necessary to evaluate the role played by the termination site and other transcriptional signals in controlling expression of the *infB* operon.

Evaluation of the amino acid sequence of IF2 reveals some remarkable features. The clearly defined periodicities in the primary structure between residues 104 and 214 (Fig. 4) suggest a regular helical conformation. The region contains a large number of polar groups, which make its penetration into a hydrophobic core unlikely but which could stabilize the helical conformation through salt linkages. Conversely, the absence of large hydrophobic side chains argues against a stable helix. It is probable that in the absence of other stabilizing interactions, the putative helical structure in this region would be unstable. Therefore, the IF2α molecule in solution may consist of two globular parts, a small NH_2-terminal domain and a large COOH-terminal domain, connected by a flexible link. It is tempting to postulate that the flexible region could assume a helical conformation as the result of an interaction of IF2 with the ribosome or fMet-

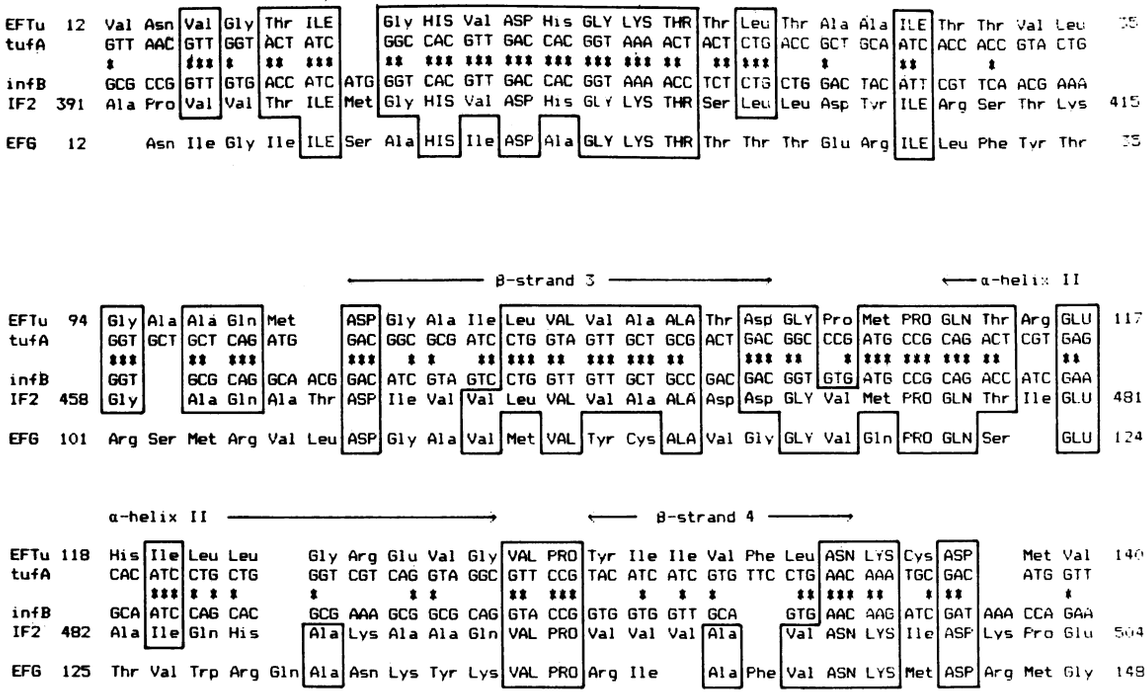

FIG. 5. Amino acid sequence homologies between IF2 and elongation factors EF-Tu and EF-G. The *infB* DNA sequence was compared to that for *tufA* (20) and two regions of extensive homology were detected: the first region (upper group) and the second region (middle and lower group). Related *infB* and *tufA* DNA sequences were aligned and stars between them signify nucleotide identities. The corresponding IF2 and EF-Tu amino acid sequences are shown, along with that for the same regions of EF-G (21). Exact amino acid homologies between IF2 and elongations factors are boxed. Amino acids identical in all three proteins are in uppercase letters. The numbers refer to the amino acid sequences. In the second region of homology, secondary structural elements of EF-Tu (22) are indicated above the amino acid sequence.

tRNA. This would lead to the drawing together of the two globular domains of IF2 and possibly to additional binding interactions with the ribosome. Crosslinking experiments have shown that IF2 is in close proximity to 16S ribosomal RNA (23), and fMet-tRNA is known to interact with the factor (3). The NH_2-terminal domain and first region of internal homology may be important in these interactions, because IF2α but not IF2β (which lacks these structures) is retained on an RNA-Sepharose column (24). Detailed chemical and physical studies of IF2 are required to elucidate how these unusual structures contribute to IF2 function.

Insight into the function of IF2 was obtained by comparing its protein and DNA structure with those of other proteins or genes involved in translation. We did not find any significant homology with initiation factors IF1 (25) and IF3 (26) nor with methionyl-tRNA synthetase (27, 28) which, like IF2, reacts with $tRNA_f^{Met}$. In contrast, amino acid sequences in the middle of IF2 show striking homology with regions of EF-Tu and to a lesser extent with those of EF-G. Both IF2 and EF-Tu bind to tRNAs, and all three factors possess GTP binding and hydrolysis activities. Thus, structural homologies between these factors are especially attractive.

One of the regions of EF-Tu (residues 94–140) with IF2 homologies contains two adjacent residues reported to be involved in GDP binding: Asn-135 and Lys-136 (22). The two residues as well as the nearby Asp-138 are conserved in IF2 and in EF-G. Thus, the Asn-498 and Lys-499 residues of IF2 may be part of the GTP binding site. The region around these residues in IF2 shows some homologies with the GTP binding proteins α- and β-tubulin and a stronger (42%) homology between IF2 (residues Gln-478 to Asp-501) and the human bladder protein p21 (29). Two residues of EF-Tu (His-66 and Cys-81) known to be involved in tRNA binding (30) are not found in regions homologous to IF2. If there are common features involved in tRNA binding, these are not yet apparent. Nevertheless, conservation of functional regions in IF2, EF-Tu, and EF-G suggests that the genes for the three factors are at least in part derived from a common ancestor.

We thank Dr. J. Nyborg for providing some basic Staden computer programs, Dr. N. J. Rasmussen for assistance in DNA sequencing, and Drs. V. I. Lim, B. F. C. Clark, A. Parmeggiani, and E. M. Bradbury for fruitful discussions. This work was supported by grants from the Centre National de la Recherche Scientifique (G.R. 18), Centre National de la Recherche Scientifique-National Science Foundation, M.R.I. (82 V 1289), Institut National de la Santé et de la Recherche Médicale (831.013), Fondation Recherche Medicale, and DuPont (to M.G.-M.) and from the American Cancer Society (NP-70) and the National Science Foundation (INT 83-12982) (to J.W.B.H.).

1. Grunberg-Manago, M., Buckingham, R. H., Cooperman, B. S. & Hershey, J. W. B. (1978) *Symp. Soc. Gen. Microbiol.* **28,** 27–110.
2. Hershey, J. W. B. (1980) *Cell Biology: A Comprehensive Treatise*, eds. Prescott, D. M. & Goldstein, L. (Academic, New York), Vol. 4, pp. 1–68.
3. Petersen, H. U., Roll, T., Grunberg-Manago, M. & Clark, B. F. C. (1979) *Biochim. Biophys. Res. Commun.* **91,** 1068–1074.
4. Plumbridge, J. A., Howe, J. G., Springer, M., Touati-Schwartz, D., Hershey, J. W. B. & Grunberg-Manago, M. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 5033–5037.
5. Ischii, S., Ihara, M., Maekawa, T., Nakamura, Y., Ushida, H. & Imamoto, F. (1984) *Nucleic Acids Res.* **12,** 3333–3342.
6. Howe, J. G. & Hershey, J. W. B. (1982) *Arch. Biochem. Biophys.* **214,** 446–451.
7. Plumbridge, J. A. & Springer, M. (1983) *J. Mol. Biol.* **167,** 227–243.
8. Clewell, D. B. (1972) *J. Bacteriol.* **110,** 667–676.
9. Maxam, A. M. & Gilbert, W. (1980) *Methods Enzymol.* **65,** 499–560.
10. Fayat, G., Mayaux, J. F., Sacerdot, C., Fromant, M., Springer, M., Grunberg-Manago, M. & Blanquet, S. (1983) *J. Mol. Biol.* **171,** 239–261.
11. Shine, J. & Dalgarno, L. (1974) *Proc. Natl. Acad. Sci. USA* **71,** 1341–1346.
12. Peacock, S., Cenatiempo, Y., Robakis, N., Brot, N. & Weissbach, H. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 4609–4612.
13. Grosjean, H. & Fiers, W. (1982) *Gene* **18,** 199–209.
14. Rosenberg, M. & Court, D. (1979) *Annu. Rev. Genet.* **13,** 319–353.
15. Kurihara, T. & Nakamura, Y. (1983) *Mol. Gen. Genet.* **190,** 189–195.
16. Grantham, R., Gautier, C., Gouy, M., Jacobzone, M. & Mercier, R. (1981) *Nucleic Acids Res.* **9,** r43–r74.
17. Staden, R. (1982) *Nucleic Acids Res.* **10,** 2951–2961.
18. Dayhoff, M. O., Barker, W. C. & Hunt, L. T. (1983) *Methods Enzymol.* **91,** 524–545.
19. Garnier, J., Osguthorpe, D. J. & Robson, B. (1978) *J. Mol. Biol.* **120,** 97–120.
20. Yokota, T., Sugisaki, H., Takanami, M. & Kaziro, Y. (1980) *Gene* **12,** 25–31.
21. Zengel, J. M., Archer, R. H. & Lindahl, L. (1984) *Nucleic Acids Res.* **12,** 2181–2192.
22. Rubin, J. R., Morikawa, K., Nyborg, J., la Cour, T. F. M., Clark, B. F. C. & Miller, D. L. (1981) *FEBS Lett.* **129,** 177–179.
23. Girshovitch, A. S., Dondon, J. & Grunberg-Manago, M. (1980) *Biochimie* **62,** 509–512.
24. Domogatckii, C. P., Vlassik, T. H. & Bezlepkina, T. A. (1979) *Proc. Acad. Sci. USSR* **248,** 240–243.
25. Pon, C. L., Wittmann-Liebold, B. & Gualerzi, C. (1979) *FEBS Lett.* **101,** 157–160.
26. Sacerdot, C., Fayat, G., Dessen, P., Springer, M., Plumbridge, J. A., Grunberg-Manago, M. & Blanquet, S. (1982) *EMBO J.* **1,** 311–315.
27. Barker, D. G., Ebel, J. P., Jakes, R. & Bruton, C. J. (1982) *Eur. J. Biochem.* **127,** 449–451.
28. Dardel, F., Fayat, G. & Blanquet, S. (1984) *J. Bacteriol.*, in press.
29. Leberman, R. & Egner, U. (1984) *EMBO J.* **3,** 339–341.
30. Clark, B. F. C., la Cour, T. F. M., Nielsen, K. M., Nyborg, J., Petersen, H. U., Siboska, G. E. & Wikman, F. P. (1984) in *Gene Expression*, Alfred Benzon Symposium 19, eds. Clark, B. F. C. & Petersen, H. U. (Munksgaard, Copenhagen), pp. 127–145.